INTEGRATING TRADITIONAL MICROBIOLOGY WITH CUTTING-EDGE (META-)GENOMICS TO ADVANCE PATHOGEN DETECTION AND TO ELUCIDATE MICROBIOME SIGNATURES OF INFECTION

A Dissertation
Presented to
The Academic Faculty

by

Angela V. Pena-Gonzalez

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biological Sciences

Georgia Institute of Technology

December, 2018

INTEGRATING TRADITIONAL MICROBIOLOGY WITH CUTTING-EDGE (META-)GENOMICS TO ADVANCE PATHOGEN DETECTION AND TO ELUCIDATE MICROBIOME SIGNATURES OF INFECTION

Approved by:

Dr. Kostas Konstantinidis, Advisor
School of Civil & Environmental Engineering
Georgia Institute of Technology

Dr. Gregory Gibson
School of Biological Sciences
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Karen Levy
Rollin School of Public Health
Emory University

Dr. Frank Stewart
School of Biological Sciences
Georgia Institute of Technology

Date Approved: November 1st, 2018

Kid, you will move mountains!

~Dr. Seuss

ACKNOWLEDGEMENTS

I would like to thank many people who have helped me throught the completition of this dissertation. During my doctoral studies, these people helped me to stay focused, stay strong and more importantly to believe in my capabilities. First, I would like to acknowledge those who guided my thinking through the process of researching. To my advisor Dr. Konstantinos T. Konstantinidis whose passion for sciences I really admire. Kostas didn't doubt in starting a new research line in his laboratory (Clinical Metagenomics) when I joined his group. Thanks for giving me the opportunity to start and develop the clinical research line and for the guidance, continuous support, patience and encouragement. To my thesis committee members –Dr. Karen Levy, Dr. King Jordan, Dr. Frank Stewart and Dr. Gregory Gibson –who have generously given their time and expertise to improve my work: thank you for your valuable contributions and advice.

I also want to thank to those who helped me get started –my labmates. To Natasha De-Leon and Shandra Justicia who were the first two 'role models' I met when I started my research at Kostas Lab. To Juliana Soto-Giron, who is the the best labmate and friend that anyone could wish for. To Luis Miguel Rodriguez, Janet K. Hatt, Carlos A. Ruiz, Minjae Kim, Despina Tsementzi and Eric Johnston: each one of you, directly or indirectly, have contributed to my education and have taught me to be a better scientitist. Thank you for your words of advice.

Last but not least, I would like to give a special thanks to my friends and family. Your love and support were critical to finish this chapter in my life. To Mom and Dad: you guys don't know how much you helped me! I could not have done it without your love and support. This is also your accomplishment. Thank you! To my son Esteban: you have been my main motor and motivation to achieve what I have accomplished so far. My main blessing in life, you are. To Miguel: Thank you for being my main emotional support during this all these years. Thanks for your patient and love during what sometimes seemed to be an endless journey. Can't wait to start new life projects together. To my brothers Camilo and Andres, and my extended family in Colombia: thanks for your constant affection and cheering. I am the first PhD in our family and I am very proud of that. I hope to have set a good example for you to follow.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANI	Average Nucleotide Identity
ANOVA	Analysis of Variance
ARGs	Antibiotic resistance genes
ATP	Adenosine Triphosphate
BoNT	Botulinum Neurotoxin
CARD	Antibiotic Resistance Database
CDC	Centers for Disease Control and Prevention
CFU	Colony Forming Unit
CI	Confidence Interval
DAEC	Diffusely -adhering *Escherichia coli*
DNA	Deoxyribonucleic Acid
dPCR	Digital PCR
EAEC	Enteroaggregative *Escherichia coli*
EcoZUR	*E. coli* en Zonas Urbanas y Rurales
EF	Edema Factor
EIEC	Enteroinvasive *Escherichia coli*
ELISA	Enzyme-linked Immunosorbent Assay
ELW	Expected likelihood weight test
EPECa	Atypical Enteropathogenic *Escherichia coli*
EPECt	Typical Enteropathogenic Escherichia coli
ETEC	Enterotoxigenic *Escherichia coli*
EYA	Egg yolk agar
FISH	Fluorescent *in situ* Hybridization Microscopy
GEMS	Global Enterics Multicenter study
gr	Grams
GS	Genome sequencer
HA	Hyaluronic acid
HGT	Horizontal Gene Transfer
IBD	Inflammatory Bowel Disease
Inc	Incompatibility groups
KH	Maximum Likelihood Kishino-Hasegawa test
LF	Lethal factor
LT	Heat-labile toxin
MAG	Metagenome-assembled genome
MCL	Markov Cluster algorithm
MiGA	Microbial Genomes Atlas
MKL	MacConkey's agar media
ML	Maximum-likelihood

MLST	Multilocus Sequencing Typing
MLVA	Multilocus variable-number tandem repeat
MSD	Moderate to severe diarrhea
ng	Nanograms
Nr	Non-redundant database
NT	Non-toxic strain
ºC	Centigrade
OGs	Orthologous genes
OTU	Operational Taxonomic Unit
PA	Protective antigen
PCA	Principal Component Analysis
pCLD	*bont/B1*-bearing plasmid
PCR	Polymerase chain reaction
pM	Picomol
pXO1	Plasmid XO1
pXO2	Plasmid XO2
qPCR	Quantitative PCR
RAST	Rapid Annotation using Subsystems
RBM	Reciprocal best matches
s.l	Sensu lato
s.s	Sensu stricto
SH	Shimodaira-Hasegawa test
SRA	Sequence Read Archive
ST	Heat-stable toxin
ST	Sequence type
STEC	Shiga toxin-producing *Escherichia coli*
TPGY	Trypticase-peptone-glucose-yeast extract
tRNA	Transfer ribonucleic acid
TTSS	Type III secretion system
UTI	Urinary tract infection
VFDB	Virulence Factors Database
WASH	Water, Sanitation and Hygiene
WGS	Whole Genome Sequencing
µl	Microliters

SUMMARY

Microbes play a central role in human health. Through normal everyday activities, the human body is exposed to countless microorganisms from the environment in addition to the hundreds of species that colonize the human body and skin. A small fraction of them are *pathogens*, which can invade and damage the human body through direct or indirect means. How pathogens perform their *in situ* activities, interact with other host-associated microorganisms and what factors control such activities remains to be fully understood. The availability of high throughput omics technologies coupled with the development of fast, powerful computational tools for the assessment of the resulting big sequencing data have provided new opportunities to better examine pathogenic microbes, not only to accurately detect them and gain basic knowledge in their biology but also to understand how they evolve and interact with other pathogenic and non-pathogenic microorganisms in a multifaceted habitat as the human body. In this thesis, a series of studies are presented that integrated clinical microbiology, epidemiology and omics techniques to study bacterial pathogens, their diversity and evolution, and to elucidate how virulent bacteria disrupt the ecology of the healthy human microbiome, especially in the intestinal tract. In this regard, Chapter 2 presents a phylogenomic study performed to resolve the true relationships between *C. botulinum* and *C. sporogenes*, two closely related species that pose differential risk for human health, and assess the frequency of horizontal transfer of the diagnostic toxin gene. Chapter 3 presents a pathogenomic study of *Bacillus anthracis* plasmids pXO1 and pXO2, executed to detect, quantify and characterize the variation among these virulent plasmids. Chapter 4 presents a phylogenomic study executed to investigate the genome diversity, population structure, virulence potential, and antibiotic resistance profile of 279 pathogenic *E. coli* strains isolated from individuals with diarrhea and controls living in urban and rural regions in Northern Ecuador. Finally, Chapter 5 presents an epidemiology and metagenomic study focused on the detection of the causative agent of diarrhea and the characterization of the signature of *Escherichia coli* infection on the gut microbiome in same cohort of young children as in Chapter 4.

CHAPTER 1

INTRODUCTION

1.1 Thesis rationale

Genomics and metagenomics can significantly enhance our understanding of pathogens and infectious diseases. With the development of high-throughput 'next generation' sequencing technologies and the technical advances to generate high-quality sequencing data, the bottleneck in implementing these omics techniques for clinical purposes has shifted from obtaining DNA sequences to post-sequencing data analysis. Although substantial progress has been already made, the widespread implementation of DNA sequencing in clinical and public health microbiology is still limited, mostly due to the lack of standardized bioinformatics methods and quality control. As sequencing technologies continue to improve and the costs continue to drop, genomic and metagenomic approaches will be increasingly applied to public health laboratories for routine diagnostics, typing and surveillance as well as the assessment of virulence factors of pathogens and disease characterization. Therefore, there is an urgent need for validation of these omics techniques in clinical specimens and standardization of high-resolution bioinformatic pipelines customized to deal with challenges associated with clinical specimens such as the co-elution of human sequences. This research effort focused on developing novel bioinformatics pipelines to process genomic and metagenomic sequencing data in order to distinguish pathogens from non-pathogenic close relatives and to elucidate the signatures of enteric pathogen infections on the gut microbiome. The specific aims of this thesis were: 1) to resolve the true phylogenetic relationships between two closely related bacterial species of medical importance, *Clostridium botulinum* and its innocuous, close relative *Clostridium sporogenes* on the basis of their core and variable genome; 2) to detect, quantify the copy number and characterize the gene variation among the virulent *Bacillus anthracis* plasmids pX1O and pXO2 in a large historical collection of globally distributed strains; 3) to investigate the genome diversity, population structure, virulence potential, and

antibiotic resistance profile of pathogenic *E. coli* strains isolated from individuals with diarrhea and controls living in urban and rural regions in Northern Ecuador and finally; 4) to characterize metagenomic signatures of the gut microbiome during *E. coli* infectious diarrhea in young children and determine whether pathotype-specific signatures exists that can be used for diagnostics of the different pathotypes.

1.2 Phylogenomics to discriminate closely related species

As whole genome sequencing (WGS) becomes more commonly used for microbial identification, it is important to evaluate the taxonomic affiliation of microbes using complete genome sequences. Phylogenomics, a concept first introduced by Jonathan Eisen in 1998 (1), was initially defined as the intersection between genomics and evolution (2, 3). In this sense, phylogenomics originally involved the reconstruction of evolutionary relationships by comparing sequences of whole genomes or at least large portions of genomes. It also included the characterization of gene family evolution, the prediction of gene functions and the detection of lateral gene transfer.

Bacterial species, particularly pathogenic ones, have been historically demarcated on the basis of a limited number of diagnostic phenotypic traits, usually virulence (e.g., toxins) or antigen factors. For example, the taxonomic classification of *Clostridium botulinum* and *Clostridium sporogenes* has been based on the production of botulinum neurotoxin (BoNT) by the former, but not the latter closely related species (4). However, this phenotypic trait (BoNT production) does not accurately mirror the true evolutionary relationships between both taxa and might pose a risk in human health. For example, non-toxic, avirulent *C. sporogenes* is usually used in industry as a surrogate organism for *C. botulinum* in the study of thermal processing for food given its metabolic similarity to toxigenic *C. botulinum* and the production of heat-resistant spores (5, 6). *C. sporogenes* is also used in academia as a research model for the study of the biology of toxigenic *C. botulinum* (7). If some strains of *C. sporogenes* could potentially acquire the BoNT via horizontal gene transfer and become highly virulent, or inversely if virulent *C. botulinum* are misidentified as *C. sporogenes*, harm and public health issues could occur.

Similarly, the taxonomy of *Bacillus cereus sensu lato* group, which includes *Bacillus cereus*, *Bacillus anthracis* and *Bacillus thuringiensis,* was initially recognized and established in early 1990s on the basis of distinct phenotypic traits (8,9). *B. anthracis* was identified as the causative agent of anthrax (10); *B. thuringiensis* was recognized as an entomopathogenic bacterium characterized by the production of parasporal crystal proteins (Cry and Cyt) (11) and finally, *B. cereus sensu stricto* was initially recognized as a common soil-dwelling microorganism that also colonized the gut of invertebrates as a symbiotic microorganism and was recently, recognized as an opportunistic human pathogen (12), but did not encode the previous virulence factors.

While the ability of bacteria to form actual discrete units (i.e., species) is a topic of intense research (reviewed in 13, 55, 56), the reliable classification of species and strains is important not only for the good practice of clinical and industrial microbiology but also for the establishment of a well-founded ecological and evolutionary framework. Phylogenetic analysis provides a way to establish a valid and biologically meaningful classification system to accurately categorize bacterial strains and define new species, especially for pathogenic microorganisms. Moreover, the horizontal transfer of genes as well as the loss of toxicity in pathogenic strains is widely recognized in bacterial species. Therefore, transitioning bacterial classification from laboratory-assessed phenotypes towards a genome-based taxonomy is critical, especially for applications in clinical microbiology and public health.

Chapter 2 provides an example of a phylogenomic pipeline developed to discriminate two closely related taxa, *C. botulinum* and *C. sporogenes* on the basis of the core genome relatedness. In particular, high-confidence clade-specific gene signatures that distinguish *C. sporogenes* from Group I *C. botulinum* were identified. These genes are not part of the taxonomic descriptions of the two species and thus, provided new reliable means to distinguish the two species. Importantly, unlike current diagnostic traits, the newly identified genes are not encoded on mobile elements or be prone to horizontal gene transfer, providing stable diagnostic traits. Such genes can be also used for rapid PCR testing. In addition, we provide examples of specific cases where we detected horizontal gene transfer of the species-diagnostic genes encoding for BoNT between both groups, which reinforce again the need to apply comparative genomics to the study of species of medical or industry importance.

1.3 Pathogenomics and evolution of virulence plasmids in *Bacillus anthracis*

Bacterial plasmids are self-replicating, extra-chromosomal, DNA elements that can play a key role in microbial population pathogenicity, adaptation and evolution due to the genes they encode (14). Plasmids act as 'vehicles' for bacterial genetic communication and promote the dissemination of a variety of traits, including virulence, enhanced fitness, resistance to antimicrobial agents, and metabolism of otherwise unusable substances (15,16). Knowledge of the relationships between plasmid features and host taxonomy is important in order to understand how the plasmids are being spread among microbes as well as the phenotypic plasticity of the strains that harbor the plasmids. In this regard, comparative genomics, defined as the genome analysis of microbes (17), can facilitate the study of plasmids encoding virulent traits.

Bacillus anthracis, the etiological agent of anthrax, typically carries two large, extra-chromosomal, double-stranded virulence plasmids called pXO1 and pXO2. Plasmid pXO1 is ~182Kb in size and carries the genes that encode for the anthrax toxin: protective antigen (PA), lethal factor (LF) and edema factor (EF). These proteins act in binary combinations to produce the two anthrax toxins: edema toxin (PA and EF) and lethal toxin (PA and LF) (18,19). Plasmid pXO2 is smaller than pXO1 (~95Kb) and harbors the genes that encode for the *cap* operon. The *cap* operon is responsible for the production of a polyglutamate capsule, which allows the pathogen to evade the host immune response by protecting itself from phagocytosis (20).

Given that plasmids pXO1 and pXO2 play a central role in the virulence of *B. anthracis*, the pathogenomic characterization of this group is usually examined in terms of the plasmids. Of particular interest is that plasmids similar to pXO1 and pXO2 have been previously found in (non-*B. anthracis*) close relatives of the *B. cereus sensu lato* group such as strains infecting primates (21-24). These findings have generated intriguing questions regarding the evolution of these self-replicating DNA elements. For example, has lateral transfer occurred between members of *B. cereus* and *B. anthracis?* If so, what is the direction of such transfers? Also, the estimated phylogenetic relationships of pathogenic strains based on the plasmid sequences are similar or perhaps dissimilar to the relationships inferred based on chromosome sequences?

Currently, *B. anthracis* plasmid detection and quantification in public health settings is accomplished by PCR amplification of diagnostic genes, namely the anthrax toxin and related genes (25). However, this approach provides low-resolution results because it targets less than 1% of the plasmid sequence and cannot provide information about the full gene content. High-throughput, sequence-based approaches can overcome these limitations. They can provide high-resolution data to not only detect and quantify plasmids based on sequencing depth but also to elucidate their full gene-content and sequence diversity. The resulting sequenced data can ultimately allow for the better understanding of the pathogenomic evolution within the group and with other close relatives.

In Chapter 3, we present a study of *Bacillus anthracis* plasmids pXO1 and pXO2 in a large collection of fully-sequenced strains that were recovered from human, animal and environmental sources around the world. The strains included in this study were part of the Zoonoses and Select Agent Laboratory's historical collection at the Centers for Disease Control and Prevention (CDC) and were acquired since the 1950s to 2013. We aimed to detect, quantify and characterize the full genomic content of the plasmids in the collection and compared the phylogenetic diversity of *B. anthracis* representatives with a large set of reference *B. cereus sensu lato* strains that included pathogenic and non-pathogenic strains carrying pXO1-like plasmids. Our results revealed that *B. anthracis* cells carry, on average, 3.86 copies of pXO1 and 2.29 copies of pXO2, and that positive linear correlations exist between the copy numbers of both plasmids. They also confirmed the remarkably stability of gene content and synteny of the plasmids and revealed in addition, no signal of plasmid exchange between *B. anthracis* and pathogenic *B. cereus* isolates but predominantly vertical descent. To the best of our knowledge, this is the largest study characterizing *B. anthracis* plasmid copy number variation and their gene content diversity using sequencing data to date and, hence, the data presented in Chapter 3 advance our understanding of the evolution and virulence potential of *B. anthracis* and its plasmids.

1.4 Genome diversity and population structure of pathogenic *E. coli* strains circulating in Northern Ecuador

E. coli is a Gram-negative, non-sporulating facultative anaerobe usually inhabiting the intestines and feces of warm-blooded animals and reptiles although it can also survive, and probably grow, in water and sediment, depending on nutrient availability and temperature (26-29). In the human gastrointestinal tract, *E. coli* is the most common facultative anaerobe microorganism, though is heavily outnumbered (~10,000 to 1) by strict anaerobic bacteria. *E. coli* strains reside in the mucus layers that covers the epithelial cells along the digestive tract and are shed into the intestinal lumen with the degraded mucus component and excreted in feces (30, 31). This bacterial group is among the first species to colonize the intestine during infancy, reaching high density before the expansion of the anaerobic component (~10^9 CFU per grams of faeces) and gradually decreasing in the elderly (32).

The relationship of *E. coli* with the host has been traditionally recognized as commensalism, in which one of the two organism benefits from the interaction between them, whereas the other is neither harmed nor helped. However, this ecological relationship is being challenged given that although it is clear that *E. coli* benefits from the hosts by getting a steady supply of nutrients (usually gluconate and low complex sugars), a stable environment, transport and dissemination, *E. coli* is also though to prevent colonization by pathogens (colonization resistance) by producing bacteriocins and other similar function compounds (33).

Paradoxically, strains (varieties) of *E. coli* can be dangerous pathogens. This microbial group has the interesting characteristic of being both a widespread commensal of vertebrates and a harmful pathogen capable of killing more than 2 million humans per year through intestinal and extra-intestinal diseases (34). Pathogenic *E. coli* strains differ from their commensal counterparts in that they have acquired distinctive traits that confer them an increased ability to invade new niches within the human body and persist. Based on virulence properties and mechanism of pathogenesis, several diarrhea-producing *E. coli* pathotypes has been recognized. These pathotypes include Enterotoxigenic *E. coli* (ETEC), Enteroinvasive *E. coli* (EIEC), Enteroaggregative *E. coli* (EAEC), Enteropathogenic *E. coli* (EPEC), Diffusely adherent *E. coli* (DAEC) and Shiga-toxin producing *E. coli* (STEC). From these, EPEC and ETEC have received special attention in developing countries as they been associated with increased diarrhea frequency in children under five years old (34, 35). To understand how *E. coli* adapts to human gut niche as a commensal microbe and what factors trigger the transition

towards virulence, it is necessary to unravel how the species is genetically structured on a global scale. Humans have distinct body sizes, diets, lifestyles, microbiotas and hygienic practices. These characteristics might substantially influence the prevalence and density of pathogenic *E. coli* populations. Therefore, the relevance of this microorganism as causative agent of disease might also be determined in part by those extrinsic forces.

In low-income countries, *E. coli* is highly prevalent in diarrheal diseases. It has been estimated that, after viral infections, Enterotoxigenic *E. coli* (ETEC) infections are responsible for more than 50% of all diarrheal deaths in children under 5 years old (35). In South America, particularly in Ecuador, few epidemiological studies have reported association between pathogenic *E. coli* and diarrhea (36-38). However, the majority of these studies have been carried out in large cities, while no study has concurrently looked at the association of *E. coli* with diarrhea in urban and rural sites over the same time period and geographic area. It is widely recognized that urbanization affects the epidemiology of emerging infectious diseases and questions regarding whether urban areas promote the establishment of more virulent *E. coli* genotypes than rural regions have been minimally explored. For example, environmental factors such as poor housing, inadequately treated drinking water supplies, sanitation and waste management in slum areas, coupled with (often uncontrolled) antibiotic usage and spatial proximity in densely populated regions may promote the establishment and rapid spread of highly virulent *E. coli* genotypes in large cities relative to rural areas.

In Chapter 4, we present a large genomic study of pathogenic *E. coli* strains circulating in urban and rural areas in Northern Ecuador as an effort to understand the extend of genome diversity and the population structure of the strains. We used molecular pathotyping and whole genome sequencing to investigate the genome diversity, pangenome structure, virulence potential, and antibiotic resistance profile of more than 200 pathogenic *E. coli* strains isolated over 17 months from individuals enrolled in a case-control study of diarrhea in four sites along an urban-rural gradient. The main findings from this study suggest, among other things, that Ecuadorian *E. coli* population is phylogenetically structured by pathotype and by host lifestyle (urban or rural) and that urban strains belonging to DAEC and atypical EPEC pathotype, carry on average, more virulence factors than rural strains.

1.5 Metagenomic-based identification of the etiological agent of infectious diarrhea and microbiome signatures caused by different *E. coli* pathotypes

The human gut microbiome, referred as the total collection of microbes (and their genomic content), that reside on and inside the digestive tract of humans, is an extremely variable, dynamic and complex habitat, containing up to 10^{12}-10^{15} microbial cells per gram of fecal material (39). The composition of this complex microbial population is highly variable over time and susceptible to both environmental and host-specific modifications (40, 41). Several studies have highlighted the critical role that the gut microorganisms play in human health. Gut microbes are involved in a variety of essential activities for the host including energy harvest and storage, fermentation and metabolism of otherwise indigestible carbohydrates, maturation of host immune-defense, and even cognitive processes (42-45).

During infectious diarrhea the structure of microbial community is often disrupted. Detailed knowledge about how the microbial community is disrupted by the abundance and the signal (e.g, excreted compounds) of the pathogen is currently limited. Much of the work to date on the gut microbiome has primarily focused on the relationship of gut community composition to chronic diseases or conditions such as obesity (46, 47), malnutrition (48), inflammatory bowel disease (IBD) (49) etc. A much more limited number of studies has examined the impacts of acute diarrhea on gut microbial communities (50, 51) using high-resolution metagenomic data.

As part of the GEM study (The Global Enteric Multicenter study sponsored by the William and Melinda Gates Foundation (52)), Pop et al., 2014 (53) used high-throughput 16SrRNA gene sequencing to study the fecal microbiota composition in children under five years of age, who had been diagnosed with moderate to severe diarrhea (MSD), and compared their results with diarrhea-free controls. This study included a total of 992 children from four low-income countries in West and East Africa (Gambia, Mali and Kenya) and Southeast Asian (Bangladesh). The main results showed that there is a significant decrease in gut microbiota diversity in diarrhea samples compared with control (which was consistently observed in all four countries), and the main differences in microbiota composition between cases and control are quantitative differences in the proportion of the most prevalent taxa.

To answer important questions regarding the burden and pathogen etiology of childhood diarrhea in developing countries, the GEM study reported the main pathogens responsible for cases of MSD in seven impoverished countries in sub-Sahara Africa and Southeast Asia. Notably, for approximately 60% of MSD in the GEMS, no known pathogen could be implicated by conventional diagnostic methods. These observations highlighted the potential presence of uncharacterized pathogens, and/or the limitations of conventional culture-dependent methods. To address some of these limitations, Huang et al., 2017 (54) reported a metagenomic study of two severe foodborne *Salmonella* outbreaks in the USA (Colorado and Alabama) with the overall goal of validating metagenomics for foodborne pathogen detection and comparing the metagenomics results to a conventional culture-based methodology. In this study, the metagenomic approach provided results consistent with the traditional culture-based approach in terms of identifying the exact pathogen/strain responsible for the disease (*Salmonella enterica* pathovar Heidelberg) but also identified cases of co-infection with *Staphylococcus aureus* and provided more precise estimates of *in-situ* pathogen abundance and intra-population diversity. In addition, this study reported an overgrowth of commensal *E. coli* as an effect of the *Salmonella* infection and shifts in the overall microbiome structure during infection relative to reference healthy samples.

Several enteric bacterial pathogens can cause diarrhea via distinctive mechanisms of pathogenicity. However, it remains undetermined whether different pathogens produce similar or perhaps, distinct alterations in the indigenous microbial community due to their characteristic mechanism of infection and/or virulence factors they produce. For example, *Escherichia coli* is usually a gut commensal of vertebrates, including humans but it can, nevertheless, cause a broad range of diseases including intestinal and extra-intestinal infections. Based on virulence factors, diarrheagenic *E. coli* have been divided into six distinct pathotypes (DAEC, EPEC, ETEC, EIEC, EAEC and EHEC), each one with preferential colonization sites, virulence mechanisms and clinical presentations.

The six *E. coli* pathotypes are characterized by distinct virulence mechanisms encoded in their respective genomes, i.e., some of them are capable of invading the cells, producing biofilms or secreting toxins. These biological differences might disturb the gut microbial community in different, distinguishable ways. We hypothesized that the distinctive infectious mechanisms used by different *E. coli* pathotypes produces

particular signatures in the sick gut microbiome. However, until now, no study has comparatively evaluated the effect of different *E. coli* pathovars on the gut microbiome during active infections. It remains unknown whether or not gut microbial signatures exist in the sick microbiome that discriminate among different *E. coli* pathotypes. For example, such signatures could include shifts in community structure and composition that can be detected and quantified using metagenomic approaches. The dogma suggests that during infection, the pathogen becomes the dominant clone in the gastrointestinal tract, overcoming the signal of the commensal population.

To address these points, In Chapter 5 we describe an integrated multi-omics approach combined with epidemiology and traditional microbiology to investigate the taxonomic composition and metabolic capability of fecal samples taken from young children in Northern Ecuador suffering infectious diarrhea relative to age-matched, healthy controls. We developed a bioinformatic approach to detect samples where *E. coli* was most likely the pathogen responsible for the disease and identified several Diffusely Adherent (DAEC), Enteropathogenic (EPEC) and Enterotoxigenic (ETEC) strain infections. These results were compared to results based on *E. coli* isolates from the same diarrhea samples that were categorized in the three pathotypes by PCR, i.e., these strains represented the etiological agents based on conventional means. Our results showed that although pathogenic *E. coli* was PCR-detected in all diarrhea samples, only about ~50% of the samples actually presented metagenomic features consistent with *E. coli* infection.

1.6 References

1. Eisen, J.A., 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome research*, *8*(3), pp.163-167.
2. Eisen, J.A. and Fraser, C.M., 2003. Phylogenomics: intersection of evolution and genomics. *Science*, *300*(5626), p.1706.
3. Delsuc, F., Brinkmann, H. and Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, *6*(5), p.361
4. Weigand, M.R., Pena-Gonzalez, A., Shirey, T.B., Broeker, R.G., Ishaq, M.K., Konstantinidis, K.T. and Raphael, B.H., 2015. Implications of genome-based discrimination between Clostridium botulinum group I and Clostridium sporogenes strains for bacterial taxonomy. *Applied and environmental microbiology*, *81*(16), pp.5420-5429.

5. Brown, J.L., Tran-Dinh, N. and Chapman, B., 2012. Clostridium sporogenes PA 3679 and its uses in the derivation of thermal processing schedules for low-acid shelf-stable foods and as a research model for proteolytic Clostridium botulinum. *Journal of food protection*, *75*(4), pp.779-792.
6. Butler III, R.R., Schill, K.M., Wang, Y. and Pombert, J.F., 2017. Genetic characterization of the exceptionally high heat resistance of the non-toxic surrogate Clostridium sporogenes PA 3679. *Frontiers in microbiology*, *8*, p.545.
7. Bradbury, M., Greenfield, P., Midgley, D., Li, D., Tran-Dinh, N., Vriesekoop, F. and Brown, J.L., 2012. Draft genome sequence of Clostridium sporogenes PA 3679, the common nontoxigenic surrogate for proteolytic Clostridium botulinum. *Journal of bacteriology*, *194*(6), pp.1631-1632.
8. Okinaka, R.T. and Keim, P., 2016. The phylogeny of Bacillus cereus sensu lato. In *The Bacterial Spore: from Molecules to Systems* (pp. 239-251). American Society of Microbiology.
9. Maughan, H. and Van der Auwera, G., 2011. Bacillus taxonomy in the genomic era finds phenotypes to be essential though often misleading. *Infection, Genetics and Evolution*, *11*(5), pp.789-797
10. Mikesell, P., Ivins, B.E., Ristroph, J.D. and Dreier, T.M., 1983. Evidence for plasmid-mediated toxin production in Bacillus anthracis. *Infection and immunity*, *39*(1), pp.371-376.
11. Aronson, A., 2002. Sporulation and δ-endotoxin synthesis by Bacillus thuringiensis. *Cellular and Molecular Life Sciences CMLS*, *59*(3), pp.417-425.
12. Økstad, O.A. and Kolstø, A.B., 2011. Genomics of Bacillus species. In *Genomics of foodborne bacterial pathogens* (pp. 29-53). Springer New York
13. Caro-Quintero, A. and Konstantinidis, K.T., 2012. Bacterial species may exist, metagenomics reveal. *Environmental microbiology*, *14*(2), pp.347-355.
14. Johnson, T.J. and Nolan, L.K., 2009. Pathogenomics of the virulence plasmids of Escherichia coli. *Microbiology and Molecular Biology Reviews*, *73*(4), pp.750-774.
15. Shintani, M., Sanchez, Z.K. and Kimbara, K., 2015. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Frontiers in microbiology*, *6*, p.242.
16. Thomas, C.M., 2014. *Horizontal gene pool: bacterial plasmids and gene spread*. CRC Press. p.52
17. Göbel, W., 2006. Bioinformatics: Data Mining Among Genome Sequences. In *Pathogenomics: Genome analysis of pathogenic microbes*. John Wiley & Sons.
18. Bergman, N.H. ed., 2011. The *Bacillus anthracis* Genome. In *Bacillus anthracis and Anthrax*. John Wiley & Sons. Chapter5, p1977
19. Okinaka, R.T., Cloud, K., Hampton, O., Hoffmaster, A.R., Hill, K.K., Keim, P., Koehler, T.M., Lamke, G., Kumano, S., Mahillon, J. and Manter, D., 1999. Sequence and organization of pXO1, the large Bacillus anthracis plasmid harboring the anthrax toxin genes. *Journal of bacteriology*, *181*(20), pp.6509-6515.
20. Candela, T., Mock, M. and Fouet, A., 2005. CapE, a 47-amino-acid peptide, is necessary for Bacillus anthracis polyglutamate capsule synthesis. *Journal of bacteriology*, *187*(22), pp.7765-7772.
21. Hoffmaster, A.R., Hill, K.K., Gee, J.E., Marston, C.K., De, B.K., Popovic, T., Sue, D., Wilkins, P.P., Avashia, S.B., Drumgoole, R. and Helma, C.H., 2006. Characterization of Bacillus cereus isolates associated with fatal pneumonias: strains are closely related to Bacillus anthracis and harbor B. anthracis virulence genes. *Journal of clinical microbiology*, *44*(9), pp.3352-3360

22. Marston, C.K., Ibrahim, H., Lee, P., Churchwell, G., Gumke, M., Stanek, D., Gee, J.E., Boyer, A.E., Gallegos-Candela, M., Barr, J.R. and Li, H., 2016. Anthrax toxin-expressing Bacillus cereus isolated from an anthrax-like eschar. *PloS one*, *11*(6), p.e0156987.
23. Antonation, K.S., Grützmacher, K., Dupke, S., Mabon, P., Zimmermann, F., Lankester, F., Peller, T., Feistner, A., Todd, A., Herbinger, I. and de Nys, H.M., 2016. Bacillus cereus biovar anthracis causing anthrax in sub-Saharan Africa—chromosomal monophyly and broad geographic distribution. *PLoS neglected tropical diseases*, *10*(9), p.e0004923.
24. Pena-Gonzalez, A., Marston, C.K., Rodriguez-R, L.M., Kolton, C.B., Garcia-Diaz, J., Theppote, A., Frace, M., Konstantinidis, K.T. and Hoffmaster, A.R., 2017. Draft genome sequence of Bacillus cereus LA2007, a human-pathogenic isolate harboring anthrax-like plasmids. *Genome announcements*, *5*(16), pp.e00181-17.
25. Riojas, M.A., Kiss, K., McKee, M.L. and Hazbón, M.H., 2015. Multiplex PCR for Species-Level Identification of Bacillus anthracis and Detection of pXO1, pXO2, and Related Plasmids. *Health security*, *13*(2), pp.122-129.
26. Tenaillon, O., Skurnik, D., Picard, B. and Denamur, E., 2010. The population genetics of commensal Escherichia coli. *Nature Reviews Microbiology*, *8*(3), p.207.
27. Gordon, D.M. and Cowling, A., 2003. The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects. *Microbiology*, *149*(12), pp.3575-3586.
28. Ishii, S., Ksoll, W.B., Hicks, R.E. and Sadowsky, M.J., 2006. Presence and growth of naturalized Escherichia coli in temperate soils from Lake Superior watersheds. *Applied and environmental microbiology*, *72*(1), pp.612-621.
29. Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M. and Konstantinidis, K.T., 2011. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences*, p.201015622.
30. Johansson, M.E., Sjövall, H. and Hansson, G.C., 2013. The gastrointestinal mucus system in health and disease. *Nature reviews Gastroenterology & hepatology*, *10*(6), p.352.
31. Rossi, E., Cimdins, A., Lüthje, P., Brauner, A., Sjöling, Å., Landini, P. and Römling, U., 2018. "It's a gut feeling"–Escherichia coli biofilm formation in the gastrointestinal tract environment. *Critical reviews in microbiology*, *44*(1), pp.1-30.
32. Penders, J., Thijs, C., Vink, C., Stelma, F.F., Snijders, B., Kummeling, I., van den Brandt, P.A. and Stobberingh, E.E., 2006. Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics*, *118*(2), pp.511-521.
33. Hudault, S., Guignot, J. and Servin, A.L., 2001. Escherichia coli strains colonising the gastrointestinal tract protect germfree mice against Salmonella Typhimurium infection. *Gut*, *49*(1), pp.47-55.
34. Lanata, C.F., Fischer-Walker, C.L., Olascoaga, A.C., Torres, C.X., Aryee, M.J. and Black, R.E., 2013. Global causes of diarrheal disease mortality in children< 5 years of age: a systematic review. *PloS one*, *8*(9), p.e72788.
35. Kotloff, K.L., Platts-Mills, J.A., Nasrin, D., Roose, A., Blackwelder, W.C. and Levine, M.M., 2017. Global burden of diarrheal diseases among children in developing countries: Incidence, etiology, and insights from new molecular diagnostic techniques. *Vaccine*, *35*(49), pp.6783-6789.
36. Riveros, M., García, W., García, C., Durand, D., Mercado, E., Ruiz, J. and Ochoa, T.J., 2017. Molecular and Phenotypic Characterization of Diarrheagenic

Escherichia coli Strains Isolated from Bacteremic Children. *The American journal of tropical medicine and hygiene*, *97*(5), pp.1329-1336.
37. Acosta, G.J., Vigo, N.I., Durand, D., Riveros, M., Arango, S., Zambruni, M. and Ochoa, T.J., 2016. Diarrheagenic Escherichia coli: prevalence and pathotype distribution in children from peruvian rural communities. *The American journal of tropical medicine and hygiene*, *95*(3), pp.574-579.
38. Eisenberg, J.N., Cevallos, W., Ponce, K., Levy, K., Bates, S.J., Scott, J.C., Hubbard, A., Vieira, N., Endara, P., Espinel, M. and Trueba, G., 2006. Environmental change and infectious disease: how new roads affect the transmission of diarrheal pathogens in rural Ecuador. *Proceedings of the National Academy of Sciences*, *103*(51), pp.19460-19465.
39. Sender, R., Fuchs, S. and Milo, R., 2016. Revised estimates for the number of human and bacteria cells in the body. *PLoS biology*, *14*(8), p.e1002533.
40. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.M. and Bertalan, M., 2011. Enterotypes of the human gut microbiome. *nature*, *473*(7346), p.174.
41. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P. and Heath, A.C., 2012. Human gut microbiome viewed across age and geography. *nature*, *486*(7402), p.222.
42. Tremaroli, V. and Bäckhed, F., 2012. Functional interactions between the gut microbiota and host metabolism. *Nature*, *489*(7415), p.242.
43. Nicholson, J.K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W. and Pettersson, S., 2012. Host-gut microbiota metabolic interactions. *Science*, p.1223813.
44. Heijtz, R.D., Wang, S., Anuar, F., Qian, Y., Björkholm, B., Samuelsson, A., Hibberd, M.L., Forssberg, H. and Pettersson, S., 2011. Normal gut microbiota modulates brain development and behavior. *Proceedings of the National Academy of Sciences*, *108*(7), pp.3047-3052.
45. Dinan, T.G., Stilling, R.M., Stanton, C. and Cryan, J.F., 2015. Collective unconscious: how gut microbes shape human behavior. *Journal of psychiatric research*, *63*, pp.1-9.
46. Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R. and Gordon, J.I., 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *nature*, *444*(7122), p.1027.
47. Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P. and Egholm, M., 2009. A core gut microbiome in obese and lean twins. *nature*, *457*(7228), p.480.
48. De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G. and Lionetti, P., 2010. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences*, *107*(33), pp.14691-14696.
49. Morgan, X.C., Tickle, T.L., Sokol, H., Gevers, D., Devaney, K.L., Ward, D.V., Reyes, J.A., Shah, S.A., LeLeiko, N., Snapper, S.B. and Bousvaros, A., 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology*, *13*(9), p.R79.
50. Bangladeshi children with acute diarrhea show fecal microbiomes with increased Streptococcus abundance, irrespective of diarrhea etiology. *Environmental microbiology*.

51. Kieser, S., Sarker, S.A., Sakwinska, O., Foata, F., Sultana, S., Khan, Z., Islam, S., Porta, N., Combremont, S., Betrisey, B. and Fournier, C., 2018. Dinleyici, E.C., Martínez-Martínez, D., Kara, A., Karbuz, A., Dalgic, N., Metin, O., Yazar, A.S., Guven, S., Kurugol, Z., Turel, O. and Kucukkoc, M., 2018. Time series analysis of the microbiota of children suffering from acute infectious diarrhea and their recovery after treatment. *Frontiers in microbiology*, *9*.
52. Kotloff, K.L., Nataro, J.P., Blackwelder, W.C., Nasrin, D., Farag, T.H., Panchalingam, S., Wu, Y., Sow, S.O., Sur, D., Breiman, R.F. and Faruque, A.S., 2013. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *The Lancet*, *382*(9888), pp.209-222.
53. Pop, M., Walker, A.W., Paulson, J., Lindsay, B., Antonio, M., Hossain, M.A., Oundo, J., Tamboura, B., Mai, V., Astrovskaya, I. and Bravo, H.C., 2014. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome biology*, *15*(6), p.R76.
54. Huang, A.D., Luo, C., Pena-Gonzalez, A., Weigand, M.R., Tarr, C. and Konstantinidis, K.T., 2016. Metagenomics of two severe foodborne outbreaks provides diagnostic signatures and signs of co-infection not attainable by traditional methods. *Applied and environmental microbiology*, pp.AEM-02577.
55. Konstantinidis, K.T., Ramette, A. and Tiedje, J.M., 2006. The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *361*(1475), pp.1929-1940.
56. Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F.L. and Swings, J., 2005. Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, *3*(9), p.733.

CHAPTER 2

GENOME-BASED DISCRIMINATION BETWEEN *CLOSTRIDUM BOTULINUM* GROUP I AND *CLOSTRODIUM SPOROGENES*: IMPLICATIONS FOR BACTERIAL TAXONOMY

Reproduced with permission from Michael R. Weigand*, Angela Pena-Gonzalez*, Timothy B. Shirey, Robin G. Broeker, Maliha K. Ishaq, Konstantinos T. Konstantinidis, and Brian H. Raphael. *Join first authors. *Appl. Environ. Microbiol.* 2015, 81(16), 5420.

2.1 Summary

Taxonomic classification of *Clostridium botulinum* is based on the production of botulinum neurotoxin (BoNT), while closely related, nontoxic organisms are classified as *Clostridium sporogenes*. However, this taxonomic organization does not accurately mirror phylogenetic relationships between these species. A phylogenetic reconstruction using 2,016 orthologous genes shared among strains of *C. botulinum* group I and *C. sporogenes* clearly separated these two species into discrete clades which showed ~93% average nucleotide identity (ANI) between them. Clustering of strains based on the presence of variable orthologs revealed 143 *C. sporogenes* clade-specific genetic signatures, a subset of which were further evaluated for their ability to correctly classify a panel of presumptive *C. sporogenes* strains by PCR. Genome sequencing of several *C. sporogenes* strains lacking these signatures confirmed that they clustered with *C. botulinum* strains in a core genome phylogenetic tree. Our analysis also identified *C. botulinum* strains that contained *C. sporogenes* clade-specific signatures and phylogenetically clustered with *C. sporogenes* strains. The genome sequences of two *bont/B2*-containing strains belonging to the *C. sporogenes* clade contained regions with similarity to a *bont*-bearing plasmid (pCLD), while two different strains belonging to the

C. botulinum clade carried *bont/B2* on the chromosome. These results indicate that *bont/B2* was likely acquired by *C. sporogenes* strains through horizontal gene transfer. The genome-based classification of these species used to identify candidate genes for the development of rapid assays for molecular identification may be applicable to additional bacterial species that are challenging with respect to their classification.

2.2 Introduction

Botulinum neurotoxins (BoNT) produce neuromuscular paralysis associated with botulism and are produced by various clostridia, most notably *Clostridium botulinum*. *C. botulinum* is a Gram-positive, anaerobic, endospore-forming bacillus that can be classified on the basis of metabolic properties into four separate groups (groups I to IV). Group I *C. botulinum* strains are proteolytic, saccharolytic, and capable of producing BoNT types A, B, and F. Group II *C. botulinum* strains are non-proteolytic and can produce BoNT types B, E, and F, while group III strains produce BoNT types C and D. Group IV strains, which are also identified as *Clostridium argentinense*, produce BoNT type G.

The nontoxic species *Clostridium sporogenes* shares nearly identical metabolic properties with group I *C. botulinum*, including the formation of lipase-positive colonies when grown on egg yolk agar. Because of these similarities and the comparable heat resistance of its spores, *C. sporogenes* has been used as a surrogate organism for *C. botulinum* in the study of thermal processing for foods (1). Previous studies have shown that some strains of *C. sporogenes* and group I *C. botulinum* can be differentiated phenotypically by soluble protein expression, measured by polyacrylamide gel electrophoresis (2) and gas-liquid chromatography of trimethylsilyl derivatives of whole-cell hydrolysates (3).

Genetic studies such as DNA-DNA hybridization and 16S rRNA gene sequence analysis have shown that these species are closely related (4–7). While acknowledging a genetic relationship at the species level, Olsen et al. (8) proposed that the species name *C. botulinum* be reserved for toxigenic strains and that *C. sporogenes* be conserved for non-toxigenic strains in order to prevent confusion between these

organisms, and this proposal has met with wide acceptance. As genome sequencing becomes more commonly used for microbial identification, it is important to evaluate this taxonomic organization in a phylogenomic context. Bacterial species are often determined on the basis of a limited number of diagnostic phenotypic traits, while the ability of bacteria to form actual discrete units (i.e., species) is a topic of intense research (9–12). Moreover, horizontal gene transfer (HGT) of genes encoding diagnostic traits may act to further reduce the boundaries between bacterial species (13).

In this study, we used genome sequence analysis to resolve *C. sporogenes* and group I *C. botulinum* strains on the basis of genetic relatedness between, as well as within, these named species. In addition, this analysis demonstrated a previously unrecognized role for HGT of some BoNT/B-encoding genes between these microorganisms.

2.3 Materials and Methods

2.3.1 Initial microbiological characterization of bacterial strains

All strains examined in this study were grown in Trypticase-peptone-glucose-yeast extract (TPGY) medium (Remel, Lenexa, KS) at 35°C under anaerobic conditions. Single colonies were isolated from egg yolk agar (EYA). Strains presumptively identified as *C. sporogenes* (Table 2.1) based on microbiological characteristics and lack of toxicity using the mouse bioassay (14) were reexamined using the botulinum toxin gene real-time PCR assay (distributed by the Laboratory Response Network, CDC) to demonstrate absence of botulinum neurotoxin genes (A to G) and by botulinum toxin enzyme-linked immunosorbent assay (ELISA) (15) to demonstrate lack of toxin (types A, B, E, and F) production. Additional toxic and nontoxic *C. botulinum* and *C. sporogenes* strains examined in this study are also described in Table 2.1.

Table 2.1. Characteristic of strains examined in this study.

Strain	Toxin Type	Species	Location	Year	Source
CDC41370	Ab	*C. botulinum*	Mexico	1996	food
CDC41370NT	NT[1]	*C. botulinum*	N/A	N/A	spontaneous mutant of CDC41370
CDC68016	B1	*C. botulinum*	West Virginia	2013	stool
CDC68016NT	NT	*C. botulinum*	N/A	N/A	spontaneous mutant of CDC68016
ATCC7949	B2	*C. botulinum*	unknown	unknown	unknown
Prevot25 NCASE	B2	*C. botulinum*	unknown	unknown	unknown
Prevot1662	B2	*C. botulinum*	unknown	unknown	unknown
ATCC51387	B2	*C. botulinum*	unknown	unknown	unknown
CDC66221	B6	*C. botulinum*	Colorado	2011	stool
11579	NT	*C. sporogenes*	unknown	1972	unknown
85-3852	NT	*C. sporogenes*	unknown	unknown	unknown
87-0535	NT	*C. sporogenes*	unknown	unknown	unknown
88-0163	NT	*C. sporogenes*	unknown	1988	blood
CDC22719	NT	*C. sporogenes*	Colorado	1977	chili sauce
CDC22720	NT	*C. sporogenes*	Colorado	1977	cheese
CDC21678	NT	*C. sporogenes*	Hawaii	1976	abdominal drainage

Table 2.1 continued

CDC21494	NT	*C. sporogenes*	Germany	1976	canned ham
CDC22679	NT	*C. sporogenes*	Puerto Rico	1977	canned meat
CDC22680	NT	*C. sporogenes*	Puerto Rico	1977	canned meat
CDC22753	NT	*C. sporogenes*	South Carolina	1977	wound (foot)
CDC22798	NT	*C. sporogenes*	South Carolina	1977	blood
CDC23091	NT	*C. sporogenes*	Oklahoma	1978	gastric fluid
CDC23284[2]	NT	*C. sporogenes*	Maryland	1978	abdominal fluid
CDC23285[2]	NT	*C. sporogenes*	Maryland	1978	peritoneal fluid
CDC24474	NT	*C. sporogenes*	Maryland	1979	abdominal fluid
CDC24545[3]	NT	*C. sporogenes*	New Zealand	1979	dotterel (bird)
CDC24442[3]	NT	*C. sporogenes*	New Zealand	1979	dotterel (bird)
CDC24533	NT	*C. sporogenes*	Arizona	1979	wound (leg)
CDC24968	NT	*C. sporogenes*	Missouri	1979	stool
CDC24726	NT	*C. sporogenes*	North Carolina	1979	liver paste
NCTC534	NT	*C. sporogenes*	Unknown	~1920	human
CDC24614	NT	*C. sporogenes*	Florida	1979	wound
CDC35120[2]	NT	*C. sporogenes*	California	1980	blood
CDC35121[2]	NT	*C. sporogenes*	California	1980	lung

Table 2.1 continued

CDC35566	NT	*C. sporogenes*	Missouri	1980	unknown
CDC35197[2]	NT	*C. sporogenes*	Alaska	1980	wound drainage
CDC35196[2]	NT	*C. sporogenes*	Alaska	1980	wound drainage

[1] NT = Non-toxic; [2] Isolates from same individual; [3] Isolates from identical culture.

2.3.2 DNA extraction and draft genome sequence assembly

Genomic DNA from all isolates was extracted from TPGY cultures as previously described (16) and purified using the DNA Clean & Concentrator-5 kit (Zymo Research, Irvine, CA) prior to library construction. Genome sequencing was performed using either the Illumina MiSeq or Life Technologies Ion Torrent PGM instrument (Table 2.2). Illumina reads were filtered and trimmed at both the 5' and 3' ends based on a Phred score threshold (*Q*) of 20 using SolexaQA (17). For each isolate, quality-trimmed reads were assembled first in parallel runs of Velvet v1.0.13 (18) and SOAP-denovo2 (19) with a range of *k*-mer values, as described previously (20). The resulting pre-contigs were pooled and then assembled using the Genome Sequencer (GS) *de novo* assembler software v2.0.01.14 (Roche, Branford, CT). Ion Torrent reads were assembled using the GS *de novo* assembler. For both platforms, assembled contigs were ordered using Mauve Contig Mover (21) with *C. botulinum* strain Loch Maree as a reference, and assembly validation was performed with Mauve Assembly Metrics (22). The resulting draft genome sequences were annotated with the Rapid Annotation using Subsystems Technology (RAST) server (23). Despite a large number of contigs in some assemblies, more than 95% of the core genes were fully recovered, suggesting that no major sequencing gaps remained.

The nucleotide sequences of *bont/B2* and *bont/B6* genes were determined by sequence read mapping using CLC bio Genomics Workbench version 7.5. For *bont/B2*-containing genomes, the ATCC 7929 *bont/B* sequence (GenBank accession number EF028395) was selected as the reference for read mapping, while the Okayama 2011

bont/B sequence (GenBank accession number AB665558) was used for *bont/B6*-containing genomes. Additional comparative analysis included alignment of selected genome sequences with pCLD (from *C. botulinum* strain Okra) using the BLAST Ring Image Generator (BRIG) (24) and alignment of the chromosomally located toxin gene cluster region of specific genomes containing *bont/B2* using the R package genoPlotR (25). The locations of specific genes within these toxin gene clusters were predicted using GeneMarkS (26).

Table 2.2. Statistics of the genome sequences generated in this study.

Strain	Platform [1]	Assembled reads	Contigs	N50	Genome Size	Cov [2]	GenBank Accession
CDC41370	PGM	536,228	980	7,449	4.1 MB	26X	LAGI01000000
CDC41370NT	PGM	695,428	254	30,380	3.8 MB	37X	LAGJ01000000
CDC68016	PGM	654,420	872	9,380	4.2 MB	33X	LAGK01000000
CDC68016NT	PGM	392,523	1714	3,800	4.3 MB	18X	LAGL01000000
ATCC7949	PGM	1,012,546	206	50,992	3.8 MB	57X	LAGE01000000
Prevot25NCASE	PGM	1,088,680	192	42,026	4.1 MB	56X	LAGN01000000
Prevot1662	PGM	775,204	593	19,719	4.9 MB	30X	LAGM01000000
ATCC51387	PGM	1,178,142	302	31,395	4.3 MB	56X	LAGD01000000
11579	PGM	577,378	317	27,263	4.4 MB	26X	JZJN01000000
85-3852	PGM	634,904	305	30,089	4.1 MB	32X	JZJO01000000
87-0535	MiSeq	1,078,369	226	46,656	3.8 MB	37X	JZJP01000000
88-0163	MiSeq	1,596,670	247	40,700	3.8 MB	53X	JZJQ01000000
CDC66221	PGM	422,086	818	9,229	4.1 MB	17X	LAGO01000000
CDC23284	PGM	580,480	506	17,646	4.1 MB	28X	LAGF01000000
CDC24442	PGM	476,812	368	14,470	3.1 MB	27X	LAGG01000000
CDC24533	PGM	793,767	344	24,303	4.1 MB	41X	LAGH01000000

[1]PGM, Life Technologies Ion Torrent Personal Genome Machine; MiSeq, Illumina. [2] Coverage

2.3.3 Core genome phylogenetic reconstruction and genetic signature identification

All predicted protein-coding genes from each of the sequenced genomes and available reference sequences were compared using an *all-versus-all* BLAST search as described previously (27). This analysis identified shared reciprocal best matches in all pairwise genome comparisons (core orthologs) of *C. sporogenes* and either all *C. botulinum* strains or only group I *C. botulinum* strains. These core orthologs were individually aligned using MUSCLE (28). The resulting alignments were concatenated to create a whole-genome alignment, and the phylogeny of genomes was reconstructed by computing a maximum-likelihood distance matrix with RAxML (29, 30).

The collection of variable genes in each genome, defined as genes absent in one or more genomes, was identified in the *all-versus-all* BLAST search described above. The presence or absence of these variable genes was used to hierarchically cluster genomes using complete linkage across a centered Pearson correlation similarity matrix using Cluster 3.0 (31). Variable genes were considered *C. sporogenes* specific if present in at least 7 out of 8 genomes and absent in all group I *C. botulinum* genomes. Conversely, genes present in 11 out of 13 group I *C. botulinum* genomes but absent in all *C. sporogenes* genomes were classified as *C. botulinum*-specific. These ratios correspond to at least 80% of the population of strains examined for each species (representing values >1 standard deviation from the mean of the population) and are therefore expected to be robust to spurious matches.

Functional annotation of predicted proteins encoded by species-specific genes was bioinformatically inferred using a combination of three independent approaches: first, sequences were searched against the UniProtKB/Swiss-Prot database (32) using DELTA-BLAST (33) with a bit score cutoff of 200. Second, conserved domains were predicted by RPS-BLAST against the position-specific score matrices of the NCBI Conserved Domain Database (34) using an e-value threshold of 0.01. Finally, a BLASTX search against the NCBI non-redundant (nr) database was performed with a bit score

cutoff of 200 and identity greater than 96%. The results of these independent searches were manually inspected, and the corresponding genes were functionally annotated based on the consensus of the three.

2.3.4 Genetic signature PCR and 16SrRNA sequencing

A total of 143 putative *C. sporogenes*-specific orthologs were identified, from which 11 candidates were selected for primer design for PCR amplification. The annotated functions and corresponding NCBI locus ID for the candidate signatures are shown in Table 2.3. PCR amplification was performed on genomic DNA extracted from 24 *C. sporogenes* isolates using primers targeting each of the 11 biomarkers (shown in Table 2.3). PCR thermo cycling conditions consisted of an initial denaturation at 95°C for 3 min followed by 35 cycles of 95°C, for 30s, 50°C for 30s, and 72°C for 90s with a final extension of 72°C for 10 min. The amplified products were examined on 1% ethidium bromide-stained agarose gels and visualized under UV trans-illumination.

A ~1.3-kb region of the 16S rRNA gene from presumptively identified *C. sporogenes* strains was amplified and sequenced using primers that were previously described (35). Sanger sequencing was performed using the GenomeLab dye termination cycle sequencing kit and a CEQ8000 genetic analysis system (Beckman Coulter, Brea, CA). Sequences were assembled and edited with Sequencher v 4.8 (Gene Codes, Ann Arbor, MI). Newly determined sequence data have been deposited in GenBank under the accession numbers shown in Table 2.2.

2.4 Results

All-versus-all BLAST comparison indicated that the pangenome for strains of *C. sporogenes* and *C. botulinum* (including groups I, II, and III) was comprised of 16,229 genes. Of these, 179 genes were shared by all strains (core orthologs), and a phylogenetic reconstruction based on their concatenated alignment revealed four distinct clades (Figure 2.1). As expected, strains from *C. botulinum* groups II, III, and I were well separated. The *C. sporogenes* strains were most closely related to a clade of group I *C. botulinum* strains, consistent with previous genomic studies (36, 37), but formed their

own discrete clade. The *C. sporogenes* clade contained *C. sporogenes* strains 88-0163, ATCC15579, 87-0535, PA3679, 85-3852, and 11579 and two *C. botulinum* type B strains, Osaka05 and Okayama2011. Furthermore, the average pairwise genome

aggregate average nucleotide identity (ANI) between genomes within the group I *C. botulinum* clade and the *C. sporogenes* clade was 93.4%, below the 95% ANI cutoff frequently used for species demarcation (38), supporting the classification of these strains into separate bacterial species. Further comparison of strains in the *C. sporogenes* and group I *C. botulinum* clades revealed a pangenome (for both clades combined) comprised of 9,306 orthologs, of which 2,016 were core and 7,290 were variable among these closely related genomes.

Hierarchical clustering based on presence or absence of the variable genes (absent in one or more genomes) among the *C. sporogenes* and *C. botulinum* strains examined in this study again separated *C. sporogenes* strains from *C. botulinum* group I, group II, and group III mirroring the phylogenetic relationships inferred from the core genome phylogeny (data not shown). When the analysis was restricted to the 16,412 variable genes present in the *C. sporogenes* and group I *C. botulinum* genomes specifically, the *C. sporogenes* strains continued to form a discrete cluster, suggesting that there exist clade-specific gene signatures in *C. sporogenes* genomes (Figure 2.2A)

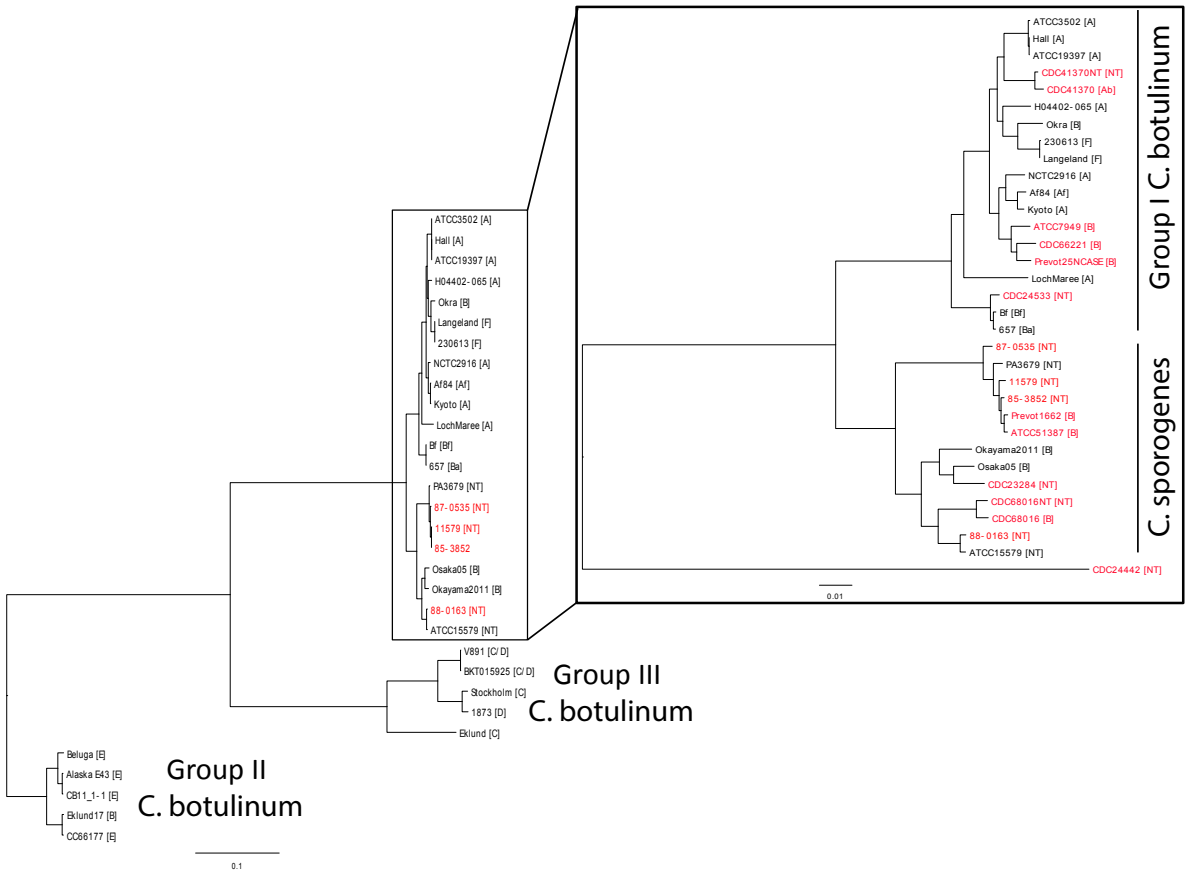


Figure 2.1. Core genome phylogenetic trees of *C. sporogenes* and *C. botulinum* genomes. The initial tree was based on the maximum-likelihood distance calculated from the alignment of 179-shared orthologous genes of C. sporogenes and C. botulinum groups II, III, and I. The tree in the inset was based on the maximum-likelihood distance calculated from the alignment of 1,451 shared orthologous genes restricted to C. sporogenes and group I C. botulinum. Specific clades discussed in the text are labeled. The BoNT produced by each strain in indicated in brackets (NT, nontoxic). Genomes sequenced in this study are in red. Additional genome sequences were retrieved from GenBank.

Selective filtering of variable orthologs identified 143 clade-specific genes that were present in at least 7 out of 8 available *C. sporogenes*-clade genomes (strains 11579, 85-3852, 88-0163, ATCC15579, 87-0535, PA3679, Osaka05, and Okyama2011) and absent in the genomes associated with the group I *C. botulinum* clade (Figure 2.2B). Of these, 65 were annotated as encoding hypothetical proteins. Annotation of the remaining genes included various functions related to nutrient uptake (e.g., heme transport and alkane sulfonate uptake) and bacterial cell defense against invasion of foreign DNA (e.g., type 1 restriction modification system). Similarly, 91 genes were observed in at least 11 out of the 13 genomes associated with the group I *C. botulinum* clade that were absent from all *C. sporogenes* strains examined. From these, 44 genes were annotated as hypothetical proteins, while the remaining genes were functionally associated with metabolic pathways (e.g., V-type ATP synthase) and cell wall modifications.

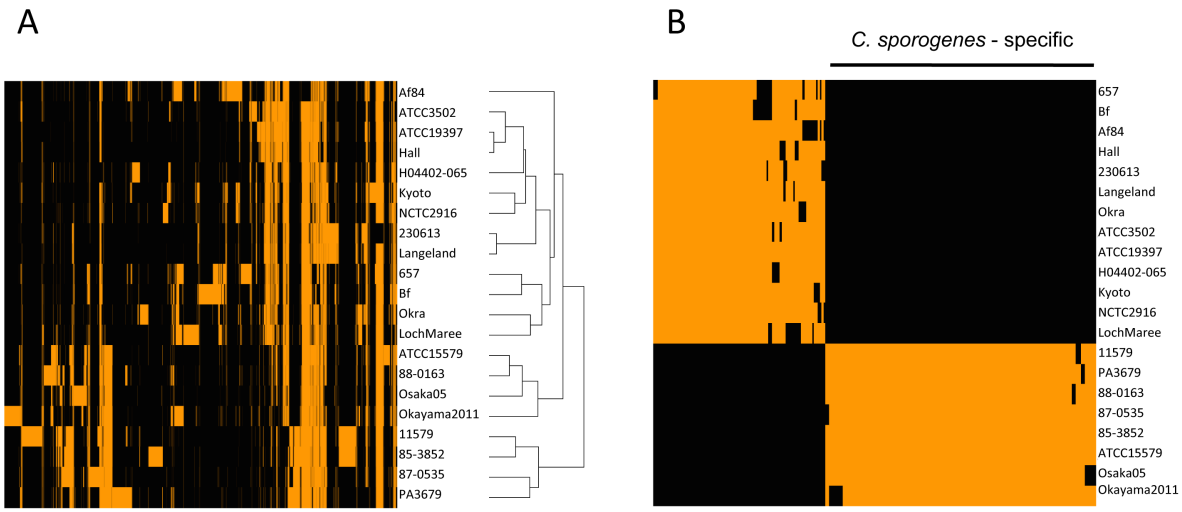


Figure 2.2. Hierarchical clustering based on the presence or absence of variable orthologs. (A) The presence (orange) or absence (black) of variable orthologs (*n* =16,412) in *C. sporogenes* and *C. botulinum* group I genomes was used to perform hierarchical clustering using a complete linkage across a centered Pearson correlation. (B) Identification of 143 orthologs genes that are present in the genomes of 7 out of 8 *C. sporogenes* strains and 91 genes that are present in 11 out of 13 group I *C. botulinum* strains.

Eleven signatures from the set of *C. sporogenes* clade-specific genomes, each representing a different operon or function, were further evaluated and annotated for potential use in PCR assays. The candidate signatures were selected from throughout the *C. sporogenes* genome in order to evenly sample the genome backbone. Manual curation of bioinformatic evidence was used to infer the putative function of each gene, and these annotation assignments are summarized in Table 2.

Table 2.3. *Clostridium sporogenes* clade-specific gene signatures. [a] The pairs of primer sequences represent forward (top) and reverse (bottom) primers.

Gene/locus	Annotated function	PCR primers *(5'→3')* [a]
hsdMSR	Type I restriction modification	TGAATCGGAAACCGATGGAC TGCCACTTGGCTTCATTTCT
ssuABC	alkanesulfonate transport	TAATCCCCTGGGCTTTACCT CCAGCTGTTATCTCATGCCA
isdAFD	heme transport	CCAGGAAAGGGCAAAAACCA TGGAGCACCAGGACACTTTA
thiHG	thiazole biosynthesis	RCGCTTGTGTCCAKCTAAAT GCCGGGTATGGAGYTAATTG
lipAS	lipoate protein ligase and synthase	TGTGACGAAGCTAATTGTCCT CTTTTTCCCCAAGACCTACCA
ltaS	lipoteichoic acid synthase	ATGGGAGAGGCCAGAAAGTA TGCCTCTCAGATACCATCCA
gadA	glutamate decarboxylase	AGATATTGGAGCGCCTGAGA CTTGTCTCCCCAATCTGGTT
dacB	D-alanyl-D-alanine carboxypeptidase	TCCAGAAGCCTGCTCTATCC AAAGTGCCTGGTGCTGCTAT
nifS	cystein desulfurase	ATAGACTCTGCTCAAACCGC GTGTAATCCGCATCTGGTCA
dapL	N-acetyl-L,L-diaminopimelate deacetylase	ACGCCTTCTGTAGCATCAAA GCCTGGACTTTAGGTACTGC

A total of 24 nontoxic lipase-positive strains presumptively identified as *C. sporogenes* were selected from the CDC culture collection to test the efficacy of *C. sporogenes* clade-specific PCR assays targeting the identified gene signatures. While PCR results were variable for some genes, a subset of five loci (*hsdMSR*, *thiHG*, *lipAS*, *ltaS*, and *dapL*) showed highly consistent results for the nontoxic strains examined

(Figure 2.3). These five conserved signatures distinguished the 24 nontoxic strains into two groups; those positive for the presence of all five signatures (referred to as “signature positive”; 18 strains) and those negative for all five signatures (“signature negative”; 6 strains).

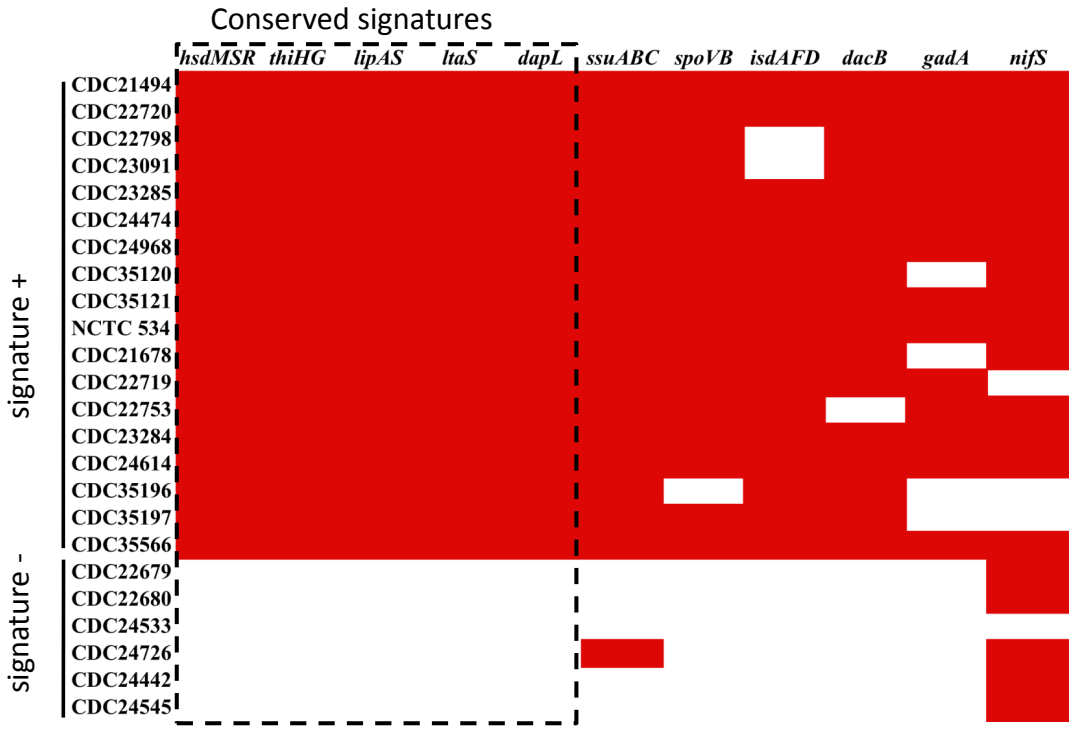


Figure 2.3. Gene signature by PCR. The results of gene signature PCR assays for each locus are shown in the columns, which are labeled across the top. The presence of a PCR product for each assay among the strains examined (indicated to the left of the figure) is indicated in red. The dotted box indicates the loci that demonstrated invariant PCR results among the strains examined.

Three putative *C. sporogenes* strains were selected for genome sequencing to determine whether the genetic signatures detected by PCR were predictive of the clade assigned by core genome phylogeny. As expected, strain CDC23284, which was signature positive, clustered with the *C. sporogenes* clade accordingly (Figure 2.1). Both

CDC24533 and CDC24442 were signature negative, but only the CDC24533 clustered with the group I *C. botulinum* clade, while strain CDC24442 formed a distinct cluster (Figure 2.1). Notably, the genome sequences of these three strains were also consistent with the 16S rRNA gene sequence analysis of all 24 strains (i.e., strains lacking the *C. sporogenes* clade-specific signatures based on PCR analysis were phylogenetically distinct from strains containing the signatures. The 16S rRNA gene sequences of strains CDC23284 and CDC24533 differed by 0.3% compared to each other and by ~1% compared to strain CDC24442. Strain CDC24442 shares 16S rRNA gene sequence identity with *C. sporogenes* subsp. *tusciae* biovar *pennavorans* (39) and clearly represents a distinct species, consistent with its distinct position in the core genome phylogeny.

Since neurotoxin gene loss should not affect core genome phylogenetic reconstruction, we also sequenced the genomes of two additional pairs of strains where a toxic and spontaneous nontoxic isolate were available. *C. botulinum* type B strain CDC68016 was isolated from the stool of an infant with botulism, and a non- toxic isolate (CDC68016NT) was also obtained during its laboratory cultivation. Both strains clustered with the *C. sporogenes* clade in the core genome phylogeny (Figure 2.1). While strain CDC68016 was initially classified as *C. botulinum* based primarily on the production of botulinum neurotoxin, this strain and its nontoxigenic progeny (CDC68016NT) are more appropriately viewed as belonging to the *C. sporogenes* clade based on their phylogenetic relationships. These findings also suggest that the type B neurotoxin gene present in strain CDC68016 was horizontally acquired, presumably from a *C. botulinum* donor. Conversely, the *C. botulinum* type Ab strain CDC41370 (isolated from a food source) and its corresponding nontoxic derivative (CDC41370NT) clustered within the group I *C. botulinum* clade.

In order to examine whether strains containing the same toxin gene variant would belong to the same genomic clade, we sequenced strain CDC66221. This strain contains the *bont/B6* variant (Figure 2.4), similar to strains Osaka05 and Okyama2011, which belong to the *C. sporogenes* clade (Figure 2.1). However, core genome phylogeny placed strain CDC66221 within the *C. botulinum* clade. Hence, strain CDC66221 can be considered a member of *C. botulinum* according to a genome-based taxonomy, suggesting that strains with the same neurotoxin gene variant do not always group within the same clade.

To better appreciate the frequency of horizontal exchange of neurotoxin genes between strains belonging to either the *C. botulinum* or *C. sporogenes* clade, we analyzed four additional *C. botulinum* type B strains (Prevot1662, ATCC 51387, Prevot25 NCASE, and ATCC 7949) containing the neurotoxin gene variant *bont/B2* (Figure 2.4). Two strains (ATCC 51387 and Prevot 1662) were positive and two strains (ATCC 7949 and Prevot 25 NCASE) were negative or the *C. sporogenes* clade-specific locus *hsdS* by PCR (data not shown). Not surprisingly, the two *hsdS*-negative strains clustered with the *C. botulinum* clade and the two *hsdS*-positive strains clustered with the *C. sporogenes* clade in the core genome phylogeny. Taken together, these data reveal that the transfer of type B neurotoxin genes between strains belonging to clades representing *C. botulinum* and *C. sporogenes* has occurred in multiple genomes and involved different *bont/B* sequence variants (e.g., strains ATCC 51387 and Prevot 1662, carrying *bont/B2*, and Osaka05 and Okyama2011, carrying *bont/B6*; all four strains group within the *C. sporogenes* clade).

Comparison of the *bont/B2*-containing genomes of *C. sporogenes* clade strains ATCC 51387 and Prevot 1662 with the *bont/B1*-bearing plasmid (pCLD) of strain Okra (Fig. 2.5A) revealed a high level of nucleotide sequence similarity. These findings are consistent with the presence of a PCR product for the previously described PL-6 plasmid marker (40) in strains ATCC 51387 and Prevot 1662 (data not shown). In contrast to these strains, which may have acquired *bont/B2* via the transfer of plasmid DNA, the *C. botulinum* clade strains ATCC 7949 and Prevot25 NCASE appear to contain *bont/B2* within their chromosomes, colocated within the *oppA- brnQ* operon (Figure 2.5B), similar to the location of the unexpressed *bont/B* found in *C.* botulinum type A(B) strain NCTC2916 (36).

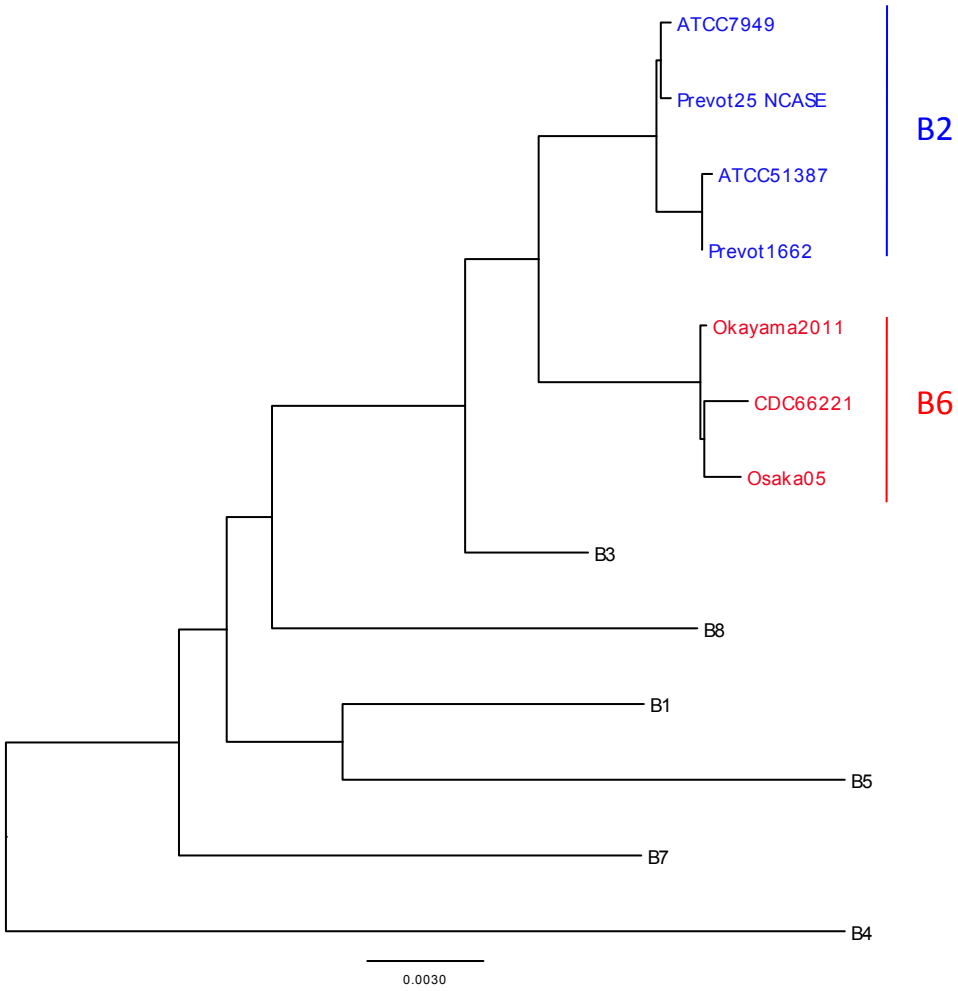


Figure 2.4. Neighbor-joining tree of *bont/B* sequences. *bont/B2* sequences are in blue, and *bont/B6* sequences are in red. Representative sequences of other *bont/B* variants were retrieved from GenBank as follows: B1, strain Okra (NC_010379); B3, strain CDC795 (EF028400); B4, strain 17B (EF051570); B5, strain CDC4013 (GU271943); B7, strain NCTC3807 (JN120760); B8, strain Chaiyaphum_2014 (KM067395).

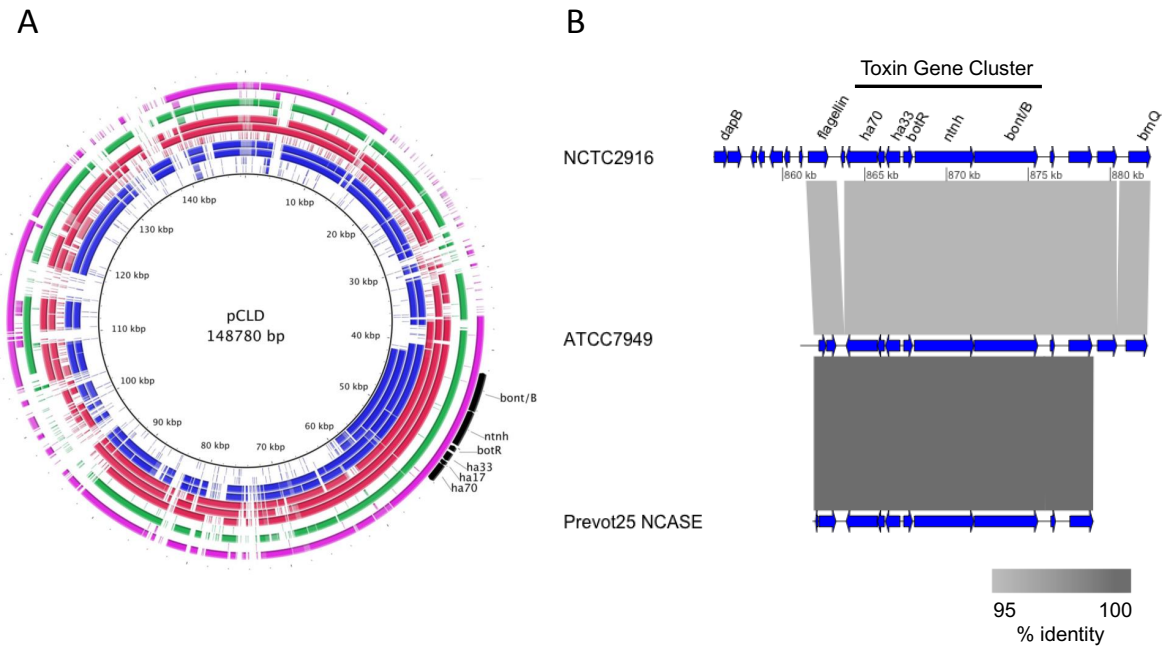


Figure 2.5. BLAST analysis of draft genome sequences. (A) Genomes were compared using BRIG with the *bont/B*-bearing plasmid (pCLD) in the *C. botulinum* type B strain Okra (GenBank accession number NC_010379). Genomes shown include (from innermost to outermost ring) ATCC 7949, Prevot 25 NCASE, ATCC 51387, Prevot 1662, Osaka05 (extrachromosomal element 2), Okayama2011, CDC66221, CDC41370NT, CDC41370, CDC68016NT, and CDC68016. Regions with >50% nucleotide identity compared to pCLD are shaded. Shading color corresponds to the *bont/B* sequence variant, as follows: blue, *bont/B2*; red, *bont/B6*; green, *bont/B5*; and purple, *bont/B1*. The positions of toxin gene cluster genes are indicated by black arrows. (B) The nucleotide sequence similarities of contigs containing the *bont/B* neurotoxin gene complex from assembled draft genome sequences ATCC 7949 and Prevot25 NCASE were compared. As a reference, the *oppA-brnQ* operon of *C. botulinum* type A(B) strain NCTC2916 (GenBank accession number NZ_ABDO02000001) was also compared to ATCC 7949. Regions sharing high nucleotide similarity are shaded.

Similarly, the genomes of *bont/B6*-containing strains Okayama2011 and CDC66221 also shared a high level of similarity with pCLD. The sequence of a 185-kb extra-chromosomal element in strain Osaka05 (GenBank accession number NZ_BA000059) also contains the *bont/B6* toxin gene cluster but shares less similar- ity with pCLD (Figure 2.5A). While all three of these *bont/B6*-containing strains appear to carry the toxin gene on a plasmid, they differ with respect to their core genome phylogeny, suggesting that such a plasmid may be able to exist in strains of either the *C. botulinum* or *C. sporogenes* clade.

2.5 Discussion

Previous studies have highlighted the challenging taxonomy of *C. botulinum* and *C. sporogenes* (7, 37). In one study, the 16S rRNA genes gene sequences belonging to 110 *Clostridium* species were compared to identify molecular markers capable of differentiating various *Clostridium* species (7). However, *C. botulinum* and *C. sporogenes* were found to cluster together in a single clade of the resulting phylogenetic tree. Additionally, comparative analysis of *in silico* restriction enzyme digestion showed that both species have exact profiles for the restriction enzymes AluI, BfaI, HaeIII, RsaI, Tru9I, and SmaI. The digestion profile for DpnII demon- strated that *C. sporogenes* was segregated into two separate populations, one of which showed exact homology with *C. botulinum* while the other was distinct from all other *Clostridium* species. These data, along with variations in DNA-DNA hybridization among different strains of *C. sporogenes* compared to *C. botulinum* (4, 5), suggest that *C. sporogenes* may be polyphyletic.

More recently, Bradbury et al. (37) reported the genome sequencing and annotation of *Clostridium sporogenes* strain PA3679, which has been used as a surrogate for group I *C. botulinum* in thermal processing of foods for over 70 years (41, 42). The percentage of shared *k*-mers (k = 25 bp) between the *C. sporogenes* PA3679 genome and several genomes of *C. botulinum* revealed that more than 85% of the *C. sporogenes* PA3679 genome matches group I *C. botulinum* strains (containing *bont/A1*). In addition, alignment of 16S rRNA gene sequences indicated a 99 to 100% nucleotide similarity between PA 3679 and several proteolytic *C. botulinum* strains, as well as other *C. sporogenes* strains.

In this study, we compared draft genome sequences to resolve *C. sporogenes* and *C. botulinum* group I strains by both core genome phylogeny and variable gene content analysis. Several *bont*-containing *C. botulinum* group I strains, such as Osaka05 and Okayama2011, were more closely related to *C. sporogenes* strains. Recent work by Kenri et al. (43) using whole-genome SNP analysis demonstrated that these two strains were distantly related to other *C. botulinum* group I strains examined. These results emphasize the conflictive nature of taxonomic designations based on the presence or absence of the botulinum neurotoxin gene, which does not reflect true phylogenetic relationships between strains.

Kalia et al. (7) suggested, "It may not be too inappropriate to suggest that *C. sporogenes* is perhaps a sub-species of *C. botulinum* or it may find its appropriate place if *C. botulinum* can be reclassified as 4 different sub-species". Our results support the latter, namely, that *C. sporogenes* is a distinct species of the genus, closely related to group I *C. botulinum*, because *C. sporogenes* strains are monophyletic on the core genome tree, show less than 95% ANI to their group I *C. botulinum* relatives but more than 95% ANI among themselves, and are characterized by specific genes and pathways that are rarely found in their relatives (Table 2.3). These genes may underlie important phenotypic differences between *C. sporogenes* and group I *C. botulinum* that remain to be elucidated. Identifying the associated phenotypes should be the subject of future work toward developing fast diagnostic tests for each of these two important bacterial taxa. While it is possible that some of the gene signatures for *C. sporogenes* may be found in other clostridial (or even non-clostridial) species, phylogenetic analysis would likely distinguish *C. sporogenes* from other species.

Among a panel of 24 nontoxic presumptive *C. sporogenes* isolated primarily from clinical samples, 18 were positive for *C. sporogenes* clade signatures. While the majority of BoNT-producing isolates examined in this study belonged to the *C. botulinum* group I clade, we observed some type B toxin producing isolates within the *C. sporogenes* clade suggesting that some toxin gene variants may be highly mobile between these two groups of organisms. Consistent with these findings, Carter et al. (44) identified only one named *C. botulinum* type B strain that grouped together with various *C. sporogenes* strains examined based on hybridization of a larger collection of strains against a ATCC 3502-specific (a group I *C. botulinum* type A strain) DNA microarray.

In a study of 63 *C. botulinum* type B strains, Franciosa et al. (45) found that more than half contained *bont/B* on a plasmid. Interestingly, neurotoxin gene-bearing plasmids were found among multiple *bont/B* sequence variants. At least some of these plasmids are likely to be mobile by conjugation, as shown by Marshall et al. (46). These findings are consistent with the results reported here that at least five *C. sporogenes*-like strains (CDC68016, ATCC 51387, Prevot 1662, Osaka05, and Okayama2011) appear to have acquired the neurotoxin gene via the horizontal transfer of plasmid DNA. Similar to another report (47), we also observed type B toxin gene loss in strains CDC41370 and CDC68016. The non-toxic strain CDC24533 belonging to the *C. botulinum* clade, specifically within a subclade containing strains that carry their toxin genes on plasmids (36, 48), presumably underwent plasmid loss. This strain was isolated from necrotic tissue of a leg wound and was not associated with a case of wound botulism, indicating that loss of toxicity by this strain probably occurred prior to infection of the wound rather than during subsequent laboratory cultivation.

Collectively, these results further support the conclusion that the current practice of classifying strains as *C. botulinum* or *C. sporogenes* based on the presence or absence of the botulinum toxin genes, respectively, cannot be highly reliable from a phylogenetic perspective, because toxin gene loss can result in confounding classification of the resulting nontoxic progeny. In this study, two strains containing *bont/B2* on a plasmid were associated with the *C. sporogenes* clade, while strains carrying this gene on the chromosome were located within the *C. botulinum* clade. Smith et al. (49) recently reported that the genome sequence of another *bont/B2*-containing strain (Prevot 594) with its neurotoxin gene on a plasmid was more related to strain Osaka05 than other toxigenic strains examined, suggesting that it was also a member of the *C. sporogenes* clade. Similarly, *in silico* multilocus sequencing typing of the draft genome sequence of *C. botulinum* strain 450 (isolated from a case of wound botulism and found to contain *bont/B2*) demonstrated a close phylogenetic relationship with *C. sporogenes* strain ATCC 15579 (50). Future work should include the study of a larger population of *bont/B2*-containing strains to determine the frequency at which *bont/B2* is plasmid borne among strains belonging to the *C. sporogenes* clade. The results presented here, which are based on a limited number of strains, indicate that the transfer of neurotoxin genes among these clades may not be a rare event.

2.6 Conclusions and recommendations

Collectively, the results presented here show that genome-derived data can offer robust resolution and classification of *C. botulinum* or *C. sporogenes* strains, even in the presence of HGT. These approaches provide a more phylogenetically accurate classification *C. botulinum* and *C. sporogenes* strains and promote a coherent bacterial species concept for these groups. Our results also underscore the limitations of the current bacterial classification system, which is frequently based on laboratory-assessed phenotypes that can be easily transferable between genetically distinct species, while providing an example of how to advance bacterial classification toward a genome-based taxonomy. This work should be applicable to additional bacterial taxa that are challenging with respect to classification.

Distinguishing *C. sporogenes* strains from *C. botulinum* isolates that have lost toxicity has important implications. Since nontoxic isolates from distinct phylogenetic clades may have different phenotypic properties, it may be useful to consider a broader range of nontoxic organisms for *C. botulinum* surrogates (especially in food protection studies) than simply those identified as *C. sporogenes*.

Genomic and metagenomic approaches for microbial identification will likely lead to fewer laboratories attempting to detect botulinum toxin in culture. The results presented here clearly suggest that genomic sequencing technologies are highly reliable and robust in correctly classifying new strains and assessing their toxin production potential (i.e., identification of neurotoxin genes). While it is difficult to determine based on traditional techniques whether a nontoxic strain with microbiological properties resembling proteolytic *C. botulinum* or *C. sporogenes* may simply be the result of toxin gene loss, the genetic signatures developed in this study can help in determining the genomic background of these nontoxic and toxic strains and also correctly interpreting previous strain classifications. The presence of *C. sporogenes* clade-specific genes would provide investigators with a high-confidence genomic signature for “true” *C. sporogenes* strains. Nontoxic strains lacking such signatures may well be nontoxic isolates of *C. botulinum* and, in the context of a botulism case investigation, may necessitate screening larger numbers of isolates for toxicity.

2.7 Acknowledgements

We thank Susan Maslanka and Carolina Lúquez for critical review of the manuscript. Genome sequencing of some strains using the MiSeq platform was performed by members of the PulseNet USA Team (CDC) and National Enteric Reference Laboratory (CDC). This work was supported by funds made available from the Centers for Disease Control and Prevention, Office of Public Health Preparedness and Response and in part by the U.S. National Science Foundation under award no. 1241046 (to K.T.K.). A.P.G. was supported by Colciencias- Colombian Administrative Department for Science, Technology and In- novation through a doctoral fellowship. T.B.S. and M.K.I. were supported by fellowships with the Oak Ridge Institute for Science and Education. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

2.8 References

1. Brown JL, Tran-Dinh N, Chapman B. 2012. *Clostridium sporogenes* PA 3679 and its uses in the derivation of thermal processing schedules for low-acid shelf-stable foods and as a research model for proteolytic *Clostridium botulin*um. J Food Prot. 75:779-92.
2. Cato EP, Hash DE, Holdeman LV, Moore WEC. 1982. Electrophoretic study of *Clostridium* species. J. Clin. Microbiol. 15:688-702.
3. Farshy DC, Moss CW. 1970. Characterization of clostridia by gas chromatography: differentiation of species by trimethylsilyl dervatives of whole-cell hydrolysates. Appl. Environ. Microbiol. 20:78-84
4. Lee WH, Riemann H. 1970. The genetic relatedness of proteolytic *Clostridium botulinum* strains. J Gen Microbiol. 64: 85-90.
5. Nakamura S, Okado I, Nakashio S, Nishida S. 1977. *Clostridium sporogenes* isolates and their relationships to *C. botulinum* on deoxyribonucleic acid reassociation. J Gen Microbiol 100:395-401.
6. Hutson RA, Thompson DE, Lawson PA, Schocken-Itturino RP, Böttger EC, Collins MD. 1993. Genetic interrelationships of proteolytic *Clostridium botulinum* types A, B, and F and other members of the *Clostridium botulinum* complex as revealed by small-subunit rRNA gene sequences. Antonie Van Leeuwenhoek. 64:273-83.
7. Kalia VC, Mukherjee T, Bhushan A, Joshi J, Shankar P, Huma N. 2011. Analysis of the unexplored features of *rrs* (16S rDNA) of the Genus *Clostridium* (2011). BMC Genomics 12:18.
8. Olsen I, Johnson JL, Moore LVH, Moore WEC. 1995. Rejection of *Clostridium putrificum* and conservation of *Clostridium botulinum* and *Clostridium sporogenes.* Int. J. Syst. Bacteriol. 45:414.

9. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 323:741-6.
10. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, Swings J. 2005. Opinion: Re-evaluating prokaryotic species. Nat Rev Microbiol. 3:733-9.
11. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 102:2567-72.
12. Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci. 361:1929-40.
13. Lawrence JG. 2002. Gene transfer in bacteria: speciation without species? Theor Popul Biol. 61:449-60.
14. Maslanka SE, Solomon HM, Sharma S, Johnson EA. 2013. *Clostridium botulinum* and its toxins. *In* Compendium of Methods for The Microbiological Examination of Foods. 5th edition. S Doores, Y. Salfinger, ML Tortorello, BW Wilcke (eds). Washington DC: American Public Health Association; 2013.
15. Maslanka SE, Lúquez C, Raphael BH, Dykes JK, Joseph LA. 2011. Utility of botulinum toxin ELISA A, B, E, F kits for clinical laboratory investigations of human botulism. Botulinum J. 2: 72–92
16. Raphael BH, Choudoir MJ, Lúquez C, Fernández R, Maslanka SE. 2010. Sequence diversity of genes encoding botulinum neurotoxin type F. Appl Environ Microbiol. 76:4805-12.
17. Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics. 11:485
18. Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821-829.
19. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaogian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 1:18.
20. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT. 2012. Individual genome assembly from complex community short-read metagenomic datasets. ISME J. *6:* 898-901.
21. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. 2009. Reordering contigs of draft genomes using the Mauve aligner. Bioinformatics 25:2071-2073.
22. Darling AE, Tritt A, Eisen JA, Facciotti MT. 2011. Mauve assembly metrics. Bioinformatics 27: 2756-2757.
23. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA,McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 9:75.
24. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 12:402.
25. Guy L, Kultima JR, Andersson SG. 2010. genoPlotR: comparative gene and genome visualization in R. Bioinformatics. 26:2334-5.
26. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nuc Acids Res. 29: 2607-2618.
27. Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the

ecology and speciation of the model bacterial species. Proc Natl Acad Sci U S A. 108: 7200-5.
28. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nuc Acids Res. 32:1792-7.
29. Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254-267.
30. Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688-90.
31. de Hoon MJ, Imoto S. Nolan J, Miyano S. 2004. Open source clustering software. Bioinformatics. 20:1453-4.
32. The UniProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). Nucl Acids Res. 42: D191-8.
33. Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. 2012. Domain enhanced lookup time accelerated BLAST. Biol Direct. 7:2.
34. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH. 2005. CDD: a Conserved Domain Database for protein classification. Nucl Acids Res. 33: D192-6.
35. Hill KK, Smith TJ, Helma CH, Ticknor LO, Foley BT, Svensson RT, Brown JL, Johnson EA, Smith LA, Okinaka RT, Jackson PJ, Marks JD. 2007. Genetic diversity among botulinum neurotoxin-producing clostridial strains. J Bacteriol. 189:818-32.
36. Hill KK, Xie G, Foley BT, Smith TJ, Munk AC, Bruce D, Smith LA, Brettin TS, Detter JC. 2009. Recombination and insertion events involving the botulinum neurotoxin complex genes in *Clostridium botulinum* types A, B, E and F and *Clostridium butyricum* type E strains. BMC Biol. 7:66.
37. Bradbury M, Greenfield P, Midgley D, Li D, Tran-Dinh N, Vriesekoop F, Brown JL. 2012. Draft genome sequence of *Clostridium sporogenes* PA 3679, the common nontoxigenic surrogate for proteolytic C*lostridium botulinum*. J. Bacteriol. 194:1631-1632.
38. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 57:81-91.
39. Ionata E, Canganella F, Bianconi G, Benno Y, Sakamoto M, Capasso A, Rossi M, La Cara F. 2008. A novel keratinase from *Clostridium sporogenes* bv. pennavorans bv. nov., a thermotolerant organism isolated from solfataric muds. Microbiol Res. 163:105-12.
40. Raphael BH, Bradshaw M, Kalb SR, Joseph LA, Lúquez C, Barr JR, Johnson EA, Maslanka SE. 2014. *Clostridium botulinum* strains producing BoNT/F4 or BoNT/F5. Appl Environ Microbiol. 80:3250-7.
41. McClung LS. 1937. Studies on anerobic bacteria. X. Heat stable and heat liable antigens in the botulinus and related groups of spore-bearing anaerobes. J. Infect. Dis. 60:122-128.
42. Townsend CT, Esty JR, Baselt FC. 1938. Heat-resistance studies on spores of putrefactive anaerobes in relation to determination of safe processes for canned foods. J. Food Sci. 3:323-346.
43. Kenri T, Sekizuka T, Yamamoto A, Iwaki M, Komiya T, Hatakeyama T, Nakajima H, Takahashi M, Kuroda M, Shibayama K. 2014. Genetic characterization and comparison of *Clostridium botulinum* isolates from botulism cases in Japan between 2006 and 2011. Appl Environ Microbiol. 80:6954-64.

44. Carter AT, Paul CJ, Mason DR, Twine SM, Alston MJ, Logan SM, Austin JW, Peck MW. 2009. Independent evolution of neurotoxin and flagellar genetic loci in proteolytic C*lostridium botulinum*. BMC Genomics. 10:115.
45. Franciosa G, Maugliani A, Scalfaro C, Aureli P. 2009. Evidence that plasmid-borne botulinum neurotoxin type B genes are widespread among Clostridium botulinum serotype B strains. PLoS One. 4:e4829.
46. Marshall KM, Bradshaw M, Johnson EA. 2010. Conjugative botulinum neurotoxin-encoding plasmids in *Clostridium botulinum*. PLoS One. 11:e11087.
47. Umeda K, Seto Y, Kohda T, Mukamoto M, Kozaki S. 2012. Stability of toxigenicity in proteolytic *Clostridium botulinum* type B upon serial passage. Microbiol Immunol. 56:338-41.
48. Marshall KM, Bradshaw M, Pellett S, Johnson EA. 2007. Plasmid encoded neurotoxin genes in *Clostridium botulinum* serotype A subtypes. Biochem Biophys Res Commun. 14:49-54.
49. Smith TJ, Hill KK, Xie G, Foley BT, Williamson CH, Foster JT, Johnson SL, Chertkov O, Teshima H, Gibbons HS, Johnsky LA, Karavis MA, Smith LA. 2015. Genomic sequences of six botulinum neurotoxin-producing strains representing three clostridial species illustrate the mobility and diversity of botulinum neurotoxin genes. Infect Genet Evol. 30:102-113.
50. Fillo S, Giordani F, Anselmo A, Fortunato A, Palozzi AM, De Santis R, Ciammaruconi A, Spagnolo F, Anniballi F, Fiore A, Auricchio B, De Medici D, Lista F. 2015. Draft Genome Sequence of *Clostridium botulinum* B2 450 Strain from Wound Botulism in a Drug User in Italy. Genome Announc. 3: e00238-15.

CHAPTER 3

GENOMIC CHARACTERIZATION AND COPY NUMBER VARIATION OF *BACILLUS ANTHRACIS* PLASMIDS PXO1 AND PXO2 IN A HISTORICAL COLLECTION OF 412 STRAINS

3.1 Summary

Bacillus anthracis plasmids pXO1 and pXO2 carry the main virulence factors responsible for anthrax. However, the extent of copy number variation within the species and how the plasmids are related to pXO1/pXO2-like plasmids in other species of the *Bacillus cereus sensu lato* group remain unclear. To gain new insights into these issues, we sequenced 412 *B. anthracis* strains representing the total phylogenetic and ecological diversity of the species. Our results revealed that *B. anthracis* genomes carried, on average, 3.86 and 2.29 copies of pXO1 and pXO2, respectively, and a positive linear correlation between the copy numbers of pXO1 and pXO2. No correlation between the plasmid copy number and phylogenetic relatedness of the strains was observed. However, strains isolated from animal tissues generally maintained a higher plasmid copy number than genomes of strains from environmental sources ($p<0.05$, *Welch two sample t-test*). Comparisons against *B. cereus* genomes carrying complete or partial pXO1 and pXO2 -like plasmids showed that plasmid-based phylogeny recapitulated that of the main chromosome, indicating limited plasmid horizontal transfer between or within these species. Comparison of gene content revealed a closed pXO1 and pXO2 pangenome, e.g., plasmids encode <8 unique genes, on average, and a

single large fragment deletion of pXO1 in one *B. anthracis* strain (2000031682) was detected. Collectively, our results provide a more complete view of the genomic diversity of *B. anthracis* plasmids, their copy number variation and the virulence potential of other *Bacillus* species carrying pXO1/pXO2-like plasmids.

3.2 Introduction

Bacillus anthracis, the etiological agent of anthrax*,* is a gram-positive endospore-forming bacterium belonging to the *Bacillus cereus sensu lato* group (1, 2). The spores are the infecting form of the bacterium and can remain viable in soils for decades as dormant, highly stable spores (2-5). *B. anthracis* has two circular, extra-chromosomal DNA plasmids named pXO1 and pXO2, which carry the major virulence factors required for pathogenesis (6). pXO1 carries the genes that encode for the anthrax toxin component: the protective antigen (PA), the lethal factor (LF) and the edema factor (EF). These proteins act in binary combinations to produce the two anthrax toxins: edema toxin (PA and EF) and lethal toxin (PA and LF) (6-8). Plasmid pXO2 harbors the genes that encode the cap operon responsible for the production of a polyglutamate capsule, which allows the pathogen to evade the host immune response by protecting itself from phagocytosis (6-8).

Given the severity of the disease and the fact that this microorganism can be easily turned into a biological weapon, it is important to characterize the diversity of the two plasmids in a large collection of strains. Currently, plasmid detection is mainly accomplished by amplification of specific markers through PCR (6, 9). Although this approach is relatively rapid, it can miss plasmids that have diverged in sequence, and cannot reveal the full gene content of plasmids. In addition, it is still not clear what is the plasmid copy number and the extent of copy number variation among members of *B. anthracis.* (8). For example, by using quantitative PCR (qPCR), Coker et al. (2003) reported ratios of up to 40.5 copies of plasmid pXO1 and 5.4 copies of plasmid pXO2 per genome (7), while Pilo et al. (2011) reported 10.89 as the average number of copies for pXO1 and 1.59 for pXO2 (10). Using digital PCR (dPCR) in three isolates, Straub et al. (2013) reported that there are likely 3-4 copies of pXO1 per cell and 1-2 copies of pXO2 (11). Sequence-based projects have also revealed that there are likely 2-3 copies

of pXO1 per chromosome copy (12). An important limitation in these previous estimates is that they have been performed in a relatively small number of isolates, which can bias the characterization of the 'true' population copy number variation. In addition, previous studies have shown that virulence of *B. anthracis* strains carrying both plasmids can differ, depending on the copy number of the plasmids (7). These results underscore the necessity to accurately quantify plasmid copy variation in a large collection of diverse *B. anthracis* isolates, and evaluate if copy variation is a phylogenetically conserved trait. High-throughput, sequence-based methods can not only detect and quantify plasmid copy number but also to elucidate gene-content and sequence diversity, which ultimately allows a better understanding of the pathogenomic evolution within the group and with other close relatives.

Further, the *B. anthracis* genome is rich in mobile elements (transposases, resolvases and integrases), which could be an important factor in plasmid gene content diversification and horizontal transfer (13). Whether or not the pXO1 and pXO2 plasmids are mobile and can be transferred between *B. anthracis* genomes as well as with other members of the *Bacillus cereus sensu lato* group, remains speculative, but it might be directly related to the virulence of the genomes and the evolutionary history of the plasmids. Gene transfer and deletion are also important for classification since strains are typically classified based on their plasmid and virulent factor content (as opposed to phylogeny) in this group.

Finally, the phylogenetic relationships within the *Bacillus cereus sensu lato* (*s.l.*) group are still problematic. *B. anthracis* belongs to the *B. cereus sensu lato* group, which includes besides *B. anthracis* two other main species: *B. cereus sensu stricto* (*s.s.*) and *B. thuringiensis* (14, 15). These species were initially recognized and established because they exhibited distinct phenotypic traits: *B. anthracis* was identified as the causative agent of anthrax (14), *B. thuringiensis* was recognized as an entomopathogenic bacterium characterized by the production of parasporal crystal proteins (Cry and Cyt), which have been widely used as a natural pesticide (15), and finally, *B. cereus s.s.* was initially recognized as a common soil-dwelling microorganism but it was later found to colonize the invertebrate guts as a symbiotic microorganism (16, 17).

DNA hybridization techniques, 16S rRNA-based typing and Multilocus Sequence Typing (MLST) schemes have progressively revealed that these species are more closely related than initially considered, and there exist limited genomic dissimilarities that allow differentiation of these species (18). This, and the fact that the main phenotypic traits for classification are carried in plasmids, has led to discussion on whether or not *B. cereus s.l.* group should be considered as a single species with characterized ecotypes and pathotypes (15, 18). Therefore, full genome analysis of newly sequenced *B. anthracis* strains and representative strains in the *B. cereus s.l.* group is critical to better elucidate the true phylogenetic relationships within the group. In addition, *B. cereus* strains encoding genetic determinants that confer pathogenic capabilities similar to those of *B. anthracis* have been previously described (19-23). Hoffmaster et al (2004, 2006) reported the isolation of *B. cereus* strains producing anthrax-like diseases in humans with clinical presentations of pneumonia and cutaneous lesions in North-America (20, 21). More recently, Antonation (2016) reported the isolation of four atypical *B. cereus* isolates (designated as *B. cereus* biovar anthracis) from death mammals (chimpanzees, gorillas, elephants and goats) in West and Central Africa (22). These isolates harbored virulence plasmids similar to those of classic *B. anthracis*.

We have also recently described the genome of *B. cereus* strain LA2007, a human-pathogenic isolate carrying a pXO1-like plasmid that showed 99.70% Average Nucleotide Identity (ANI) with *B. anthracis* Ames pXO1 (24) (Appendix A.1). Interestingly, the pXO1-like plasmids of pathogenic *B. cereus* strains reported to date are similar but not identical to those found in *B. anthracis*. Therefore, determining how diverse are the anthrax-like plasmids in terms of genetic backbone, modularity and phylogeny is critical not to only develop more accurate detection tools but also to understand the pathogenomic evolution of virulence determinants within the *B. cereus s.l.* group.

In this study we used next generation sequencing data to detect, quantify and characterize the full genomic content of *B. anthracis* plasmids pXO1 and pXO2 in a collection of 412 newly sequenced strains that represent well the global diversity of the species recovered to date. We also compared the phylogenetic diversity of *B. anthracis* representatives with a set of 106 *B. cereus sensu lato* strains that included pathogenic and non-pathogenic strains carrying anthrax-like plasmids.

3.3 Materials and Methods

3.3.1 Collection description

The collection of genomes analyzed in this study is part of the Zoonoses and Select Agent Laboratory's historical strain collection at the Centers for Disease Control and Prevention. The strains included in the study were acquired from human, animal, and environmental sources worldwide from the 1950s to 2013. The complete set has been deposited in the NCBI Sequence Read Archive (SRA) under BioProject ID 264742

3.3.2 Growth conditions, DNA extraction and sequencing

DNA from isolates was extracted using QIAamp DNA Blood Mini Kit (Qiagen, Valencia, CA) or Maxwell 16 Instrument (Promega, Madison, WI). For the QIAamp extraction, cells were grown overnight in heart infusion broth (Remel, Lenexa, KS). Cells were pelleted by centrifugation for 10 min at 5,000 x g. Broth was removed and DNA was extracted using the Qiagen QIAamp DNA Blood Mini Kit following manufacturer's protocol for isolating gram-positive bacteria. For DNA extractions on the Maxwell instrument, cells were grown overnight on trypticase soy agar with 5% sheep blood. Cells were mechanically disrupted by vortexing for 2 min in a suspension of silica beads and TE buffer. The suspension was centrifuged for 30 seconds at 10,000 x g. 300 µL of the resulting supernatant was used for DNA extraction following manufacturer's protocol for blood and cells. Sequencing was performed on an Illumina GAIIx using TruSeq chemistry.

3.3.3 Read quality control, assembly, and gene prediction

Raw reads were initially screened for adaptor sequences using Scythe (32) and trimmed at both 5' and 3' ends based on a PHRED score cutoff of 20 using SolexaQA++ (33). Reads < 50 bp after trimming were discarded. Quality-filtered reads were *de novo* assembled using IDBA-UD with pre-corrections (34) and the percent of contamination and genome completeness were assessed based on either recovery of lineage-specific marker genes using CheckM (35) or recovery of essential genes (single copy) in bacterial and archaeal genomes using the script *HMM.essential.rb* available at

Enveomics collection (36). Protein-coding sequences were predicted using MetaGeneMark (37), and 16S rRNA gene sequences were identified using barrnap 0.6 (https://github.com/tseemann/barrnap). All predicted genes from the assemblies were taxonomically annotated using MyTaxa (38) and the taxonomic distributions of adjacent genes (in windows of 10 genes) in the concatenated assembly were inspected for possible contamination through barplots. The above-described methods and scripts for read quality control, assembly and gene prediction were used as part of MiGA (Microbial Genomes Atlas), a system developed in our laboratory for data management and processing of microbial genomes and metagenomes (http://microbial-genomes.org/).

3.3.4 *B. anthracis* and *B. cereus* s.l. reference genomes

Assembled sequencing data for 36 additional *B. anthracis* and raw sequencing reads for 130 *B. cereus sensu lato* reference strains were downloaded from the nucleotide database or the Sequencing Read Archive (SRA) at NCBI (http://www.ncbi.nlm.nih.gov/sra) with the accession numbers described in Supplemental Table 2. Reference strains were processed in parallel with CDC *B. anthracis* collection as described above. After quality control inspection, 26 *B. cereus* reference strains showing ≥ 20% contamination as calculated with CheckM (see above) were excluded from the analysis.

3.3.5 Plasmid copy number estimation

Whole genome sequencing enabled us to estimate the copy number for each plasmid relative to the chromosome copies in each sequence library. Copy number was estimated as the ratio of the average number of reads mapping to the plasmid and the chromosome with ≥ 95% nucleotide identity. To speed up computational processing, read sets were randomly subsampled to a level where conclusions would not change. We varied library sizes and calculated the pXO1 copy number for three libraries of different sizes (large, medium and small), and as low as 10% of the library size did not have an effect in copy number estimation. Quality-filtered sequence libraries were therefore subsampled to 10% of their size and blastn mapped to three targets: the reference *B. anthracis* Ames ancestor (GCF_000008445.1) plasmids pXO1 (NC_007322.2) and pXO2 (NC_007323.2) and each assembled genome. Read depths were calculated for each library using the function '*enve.recplot*' incorporated in the R

package '*enveomics.R'* (36). Using the same R function, read recruitment plots were generated per each library to quantify and visualize the coverage across the full length of the reference plasmids and determine the presence/absence of the entire plasmid. Presence was considered true if the calculated average sequencing depth across the full reference was ≥ 2X in the subsampled library.

3.3.6 Average Nucleotide Identity (ANI) distances and medoids

The Average Nucleotide Identity (ANI) (39, 40) between the set of genomes were calculated using the command line interface of MiGA (Microbial Genome Atlas, https://github.com/bio-miga/miga). Briefly, MiGA calculated a matrix of distances with 1-ANI for all pairs of genomes considered in the database. After this, clusters in the matrix were identified using the PAM algorithm (Partitioning Around the Medoids) (41) with *k* medoids, where *k* was determined by the local gain in the average Silhouette width (42) for each level of clustering until a group of five or fewer genomes was reached. Here, medoids are representative strains in the diversity space. After this, a dendrogram was built based on ANI distances (1-ANI) using hierarchical agglomerative clustering with the Ward criterion (43).

3.3.7 Phylogenetic signal in plasmid copy number

Phylogenetic conservatism of plasmid copy number was determined through the calculation of the Blomberg's *K* statistic (44) included in the function '*phylosignal*' of the R package '*Picante*' (45). *K* values of 1 correspond to a Brownian motion process, which implies some degree of phylogenetic signal or conservatism. *K* values closer to zero correspond to a random or convergent pattern of evolution, while *K* values greater than 1 indicate strong phylogenetic signal and conservatism of traits.

3.3.8 Phylogenomic relationship of plasmids and chromosome based on ANI

Large contigs (≥ 500bps) with ≥ 80% identity and ≥ 80% query coverage to either pXO1 or pXO2 *B. anthracis* Ames ancestor reference sequences were considered to be pXO1 or pXO2 homologous and were extracted from the assemblies. Dendrograms based on ANI distances were built for the plasmids and chromosomes as described previously and subsequently compared through tanglegrams using the R package *Dendextend* version 1.2.0 (46). Statistical correlation between pairs of dendrograms was evaluated with two

parameters: Baker's Ω index correlation (47) and the cophenetic distant correlation (48), both of them included in the R package *Dendextend*.

3.3.9 Read-based genomic gene content analysis

pXO1 and pXO2 orthologs genes among *B. anthracis* genomes were identified using a reciprocal-best match (RBMs) blastn approach as described in Weigand (2015) (49). In brief, the sequences of the predicted genes in the plasmid sequence of one strain were searched against the predicted genes of all remaining set of strains in a pair-wise fashion using blastn (50). Reciprocal best matches (RBMs) were identified when the best match was bidirectional for the pair of strains being compared and there was at least 70% nucleotide identify and 70% query gene coverage using rbm.rb (36). Next, orthologous groups (OGs) in reciprocal best matches were identified using the unsupervised Markov Cluster algorithm (MCL) implemented in *ogs.mcl.rb* in the Enveomics collection (36) with default settings: 1.5 inflation parameter and Bit-score as parameter to weight edges. Descriptive statistics on the set of orthology groups were estimated using the script *ogs.stats.rb* (Enveomics collection). Genes conserved in all genomes were denoted as core orthologous genes. Genes conserved in some but not all of the strains were identified as variable orthologous genes. Representative orthologous genes from the previous analysis (including both core and variable genes) were randomly selected and extracted to generate a pangenome or 'bag of genes'. To better determine the presence/absence of the genes included in the pangenome, we recruited raw sequencing reads against the predicted genes on the plasmid. For this, FastA libraries were subsampled to 500,000 reads per sample and mapped against the set of representative orthologous genes using blastn. The maximum number of target sequences in the database was set to 1 (best match). After this, the observed and estimated sequencing depth as well as the number of reads mapping to each gene in the database was calculated using the script '*BlastTab.seqdepth_ZIP.pl*' from the enveomics collection (36) assuming a Zero-Inflated Poisson Distribution to correct for non-covered positions with parameters estimated as described in (Beckett, 2014) (51). Orthologous genes with zero-inflation ≥ 0.3, which represent the fraction of the gene that is not covered, were excluded. In other words, only genes with ≥ 70% coverage were considered to be present. The calculated average sequencing depth for the genes in pXO1 was 32.2X and for the genes in pXO2 was 19.9X. To determine the copy number

of the genes in each plasmid, the sequencing depth calculated for each gene in each strain was normalized by the median sequencing depth of each strain and reported through a dendrogram of hierarchical clustering.

3.3.10 Genomic characterization of *B. anthracis* and *B. cereus* strains carrying pXO1 and/or pXO2-like plasmids

Orthologous genes among *B. anthracis* medoids and *B. cereus* genomes carrying anthrax-like plasmids were identified through the reciprocal-best match (RBMs) blastn approach as described above. Core genes were extracted and aligned to estimate phylogenetic relationships between *B. cereus* and *B. anthracis*. The set of core genes were filtered to remove in-paralogous genes and aligned using MUSCLE v3.8.31 (52) with default parameters. The aligned outputs were saved in FastA format and the script *Aln.cat.rb* from the Enveomics collection was used to concatenate the multiple alignments into a single file and to remove invariable sites, defined as columns with only one state and undefined characters. Phylogenetic reconstructions were performed with either RAxML version 8.1.21 (53) or or FastTree version 2.1.7 (60) with the GTR model for nucleotides in both cases. On the other hand, the collection of variable genes, defined as genes absent in 1 or more genomes, were identified as described above and the presence or absence of these variable genes was used to cluster genomes hierarchically using a complete linkage across a centered Pearson correlation similarity using the function '*heatmap2*' contained in the R package gplots v3.0.1 (https://CRAN.R-project.org/package=gplots). Functional annotation of variables genes was bioinformatically inferred through a BLASTp search against the RefSeq protein and the UniProtKB/Swiss-Prot databases with percent identity greater than 45% and minimal query coverage of 70%.

Tests for incongruence between phylogenetic trees were calculated using Tree Puzzle 5.2 and Maximum Likelihood (ML) (55). ML analysis was carried out using empirically derived base frequencies, transitions to transversions ratio estimated from dataset, the HKY model of substitution, a gamma distribution model for site rate variation with α-parameter estimated from dataset and four (4) gamma rate categories. Topological variability or distances among trees derived from individual orthologous

genes (OGs) were calculated using the Kendall and Colijn metric (61) implemented in R package 'treeSpace' (62). Tanglegram entanglements were calculated as described in (46).

3.4 Results

3.4.1 Estimated plasmid copy number and covariance

In this study, a total of 412 *B. anthracis* strains were newly sequenced. The whole-genome comparison of these genomes will be reported elsewhere; here, we focused our analyses on the plasmid sequences. Libraries had an average sequencing depth of 135.4X with a median value of 128.3X and a minimum value of 9.8X. To estimate pXO1 and pXO2 copy number, we calculated the ratio of plasmid sequence depth (using *B. anthracis* Ames ancestor plasmid sequences as references to recruit reads) to the average sequencing depth for the chromosome. We identified a total of 58 and 42 strains that completely lacked pXO1 and pXO2, respectively, or had too few reads (i.e., <2X sequencing depth after subsampling; see Materials and Methods for details) mapping on the plasmid (i.e., 42 and 62 strains for pXO1 and/or pXO2, respectively), and therefore were not included in the estimations.

By calculating the ratio of plasmid to chromosome sequencing depth, we estimated that *B. anthracis* cells maintain on average 3.86 copies of plasmid pXO1 (Stdev =1.27) and 2.29 copies of pXO2 (Stdev =0.54) indicating a general pattern in which there are almost twice (1.68 times) as many copies of pXO1 relative to pXO2 (Figure 3.1A). In addition, we observed a large variation in copy number where some strains carried up to 7.8 copies of pXO1, contrasting with pXO2, where the maximum number of copies was 4.5 and it was generally less variable. We also observed a considerable degree of positive linear correlation between the copy numbers of pXO1 vs. that of pXO2 (Pearson's $r = 0.68$, Spearman's $\rho = 0.62$) **(**Figure 3.1B**).**

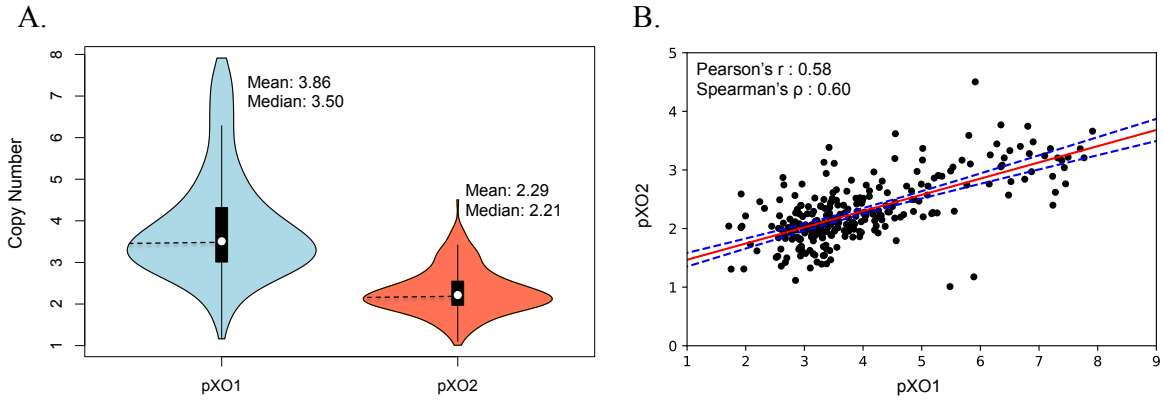


Figure 3.1. Copy number estimation of *Bacillus anthracis* plasmids pXO1 and pXO2. (A) Plasmid copy number distribution calculated for strains carrying one or both plasmids. We estimated that *B. anthracis* cells maintain in average 3.86 copies of plasmid pXO1 and 2.29 copies of pXO2 indicating a general pattern in which there are almost twice (1.68 times) as many the number of pXO1 than pXO2. (B) Correlation analysis between pXO1 and pXO2 estimated copy number showed a high degree of linear positive correlation (*Pearson's r = 0.68, Spearman's* ρ*=0.62*). Red line shows the estimated linear regression model and the dashed blue lines depict the upper and lower confidence intervals at 95%.

3.4.2 Plasmid copy number variation and genomic relatedness

Next, we evaluated whether plasmid copy number in *B. anthracis* is a phylogenetically conserved trait. To test this hypothesis, we calculated the Blomberg's *K* statistics (43), which relates the amount of phylogenetic signal to expectation under Brownian motion of character evolution using a dendrogram derived from an ANI distance tree (Figure 3.2). We observed no correlation between the plasmid copy number and genome-average nucleotide identity (ANI) distances among strains (Blomberg's *K = 0.013,* for pXO1 and *K = 0.014,* for pXO2; and Figure 3.2), indicating that plasmid copy number is not phylogenetically conserved. Thus, closely related strains do not necessarily carry similar plasmid copy numbers.

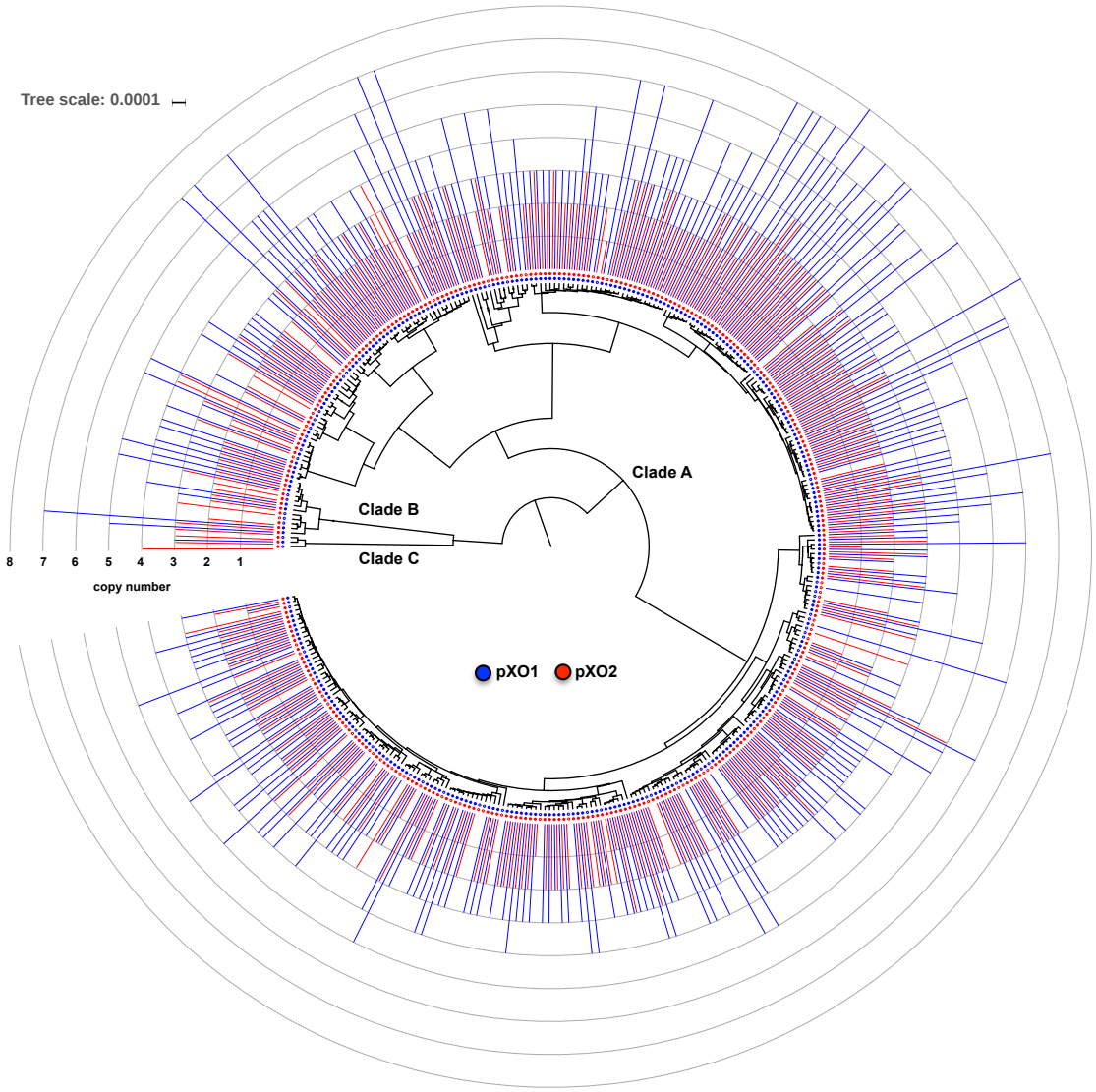
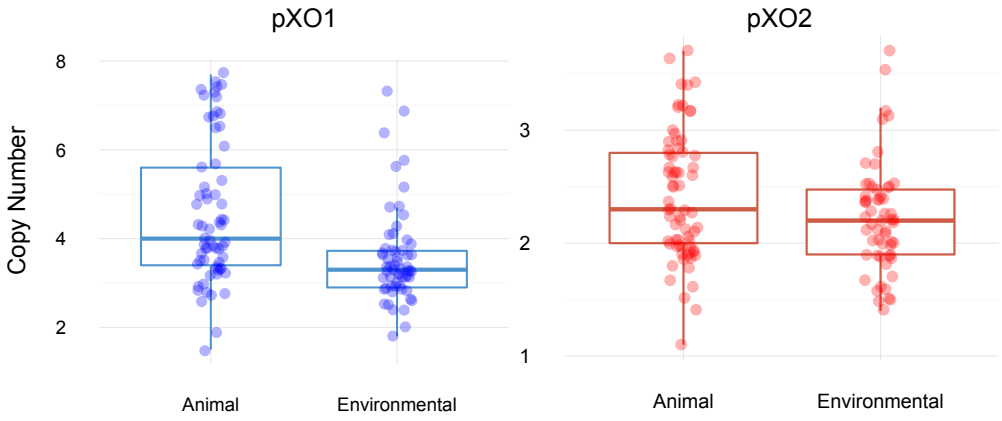


Figure 3.2. Lack of phylogenetic conservatism of *Bacillus anthracis* plasmid copy number. Dendrogram was constructed based on the Average Nucleotide Identity (ANI) distances calculated for 412 *B. anthracis* strains. Presence or absence of pXO1 (inner circle in blue) and pXO2 (inner circle in red) and estimated plasmid copy number (barplots) are shown. Strains with high and low plasmid copy numbers were found to be dispersed across the three main clades A (397), B (12) and C (3), and no apparent clusters were evident. The tree scale corresponds to 1-ANI distance.

3.4.3 Plasmid copy number depends on source of isolation

To evaluate any correlation existing between the estimated plasmid copy number and the source of the strains, a comparative analysis was performed on 127 strains for which biological source information was available. Biological source was defined as human, animal or environmental based on the sample from which each strain was isolated. Results showed that *B. anthracis* isolated from human and animal sources generally maintained a higher plasmid copy number than strains isolated from environmental sources ($p = 9.7e\text{-}5$ for pXO1 and $p = 0.05$ for pXO2, Welch two sample t-test) (Figure 3.3).

To exclude the possibility that this observation was the result of DNA extraction method, given that two protocols were implemented (Qiagen and Promega; see Methods and Materials for further details), we performed a two-sample t-test analysis comparing plasmid copy number between the two extraction methods. The results reveled no significant difference ($p = 0.11$ for pXO1; $p = 0.81$ for pXO2). In addition, we performed an analysis of variance (ANOVA) to determine the influence of DNA extraction method and biological source (two independent variables) in explaining the values of plasmid copy number (the continuous dependent variable). The results showed that the variation explained by the biological source was significant ($F = 6.23$, $p = 0.01$) while the variation explained by the extraction method was not significant ($F = 0.072$, $p = 0.7$).



	pXO1				pXO2			
	N	Mean	SD	*p*-value	N	Mean	SD	*p*-value
Total	125				127			
Animal	65	4.55	1.59		69	2.41	0.56	
Environmental	60	3.58	1.07	9.7e-05***	58	2.22	0.47	0.05*

Figure 3.3. Plasmid copy number depends on source of isolation. *B. anthracis* isolates obtained from animal sources (both human and animal) maintain, in general, a higher plasmid copy number than strains isolated from environmental sources. Top: distribution of pXO1 (left) and pXO2 (right) copy number for each group of genomes. Boxplots represent the first and third quartile and the horizontal segment represents the median value. Bottom: table with the statistic of the copy number distribution observed. Note that the average number of pXO1 plasmids in strains from animal origin was significantly higher than the mean copies estimated for strains obtained from environmental samples (*p*-values were calculated based on parametric *Welch's two sample t-test*).

3.4.4 Plasmid- vs. chromosome-based phylogenetic relationships

To determine whether plasmid-based ANI clustering resembled that of the chromosome, we analyzed strains for which plasmid pXO1 and/or pXO2 were detected, in addition to 36 *B. anthracis* reference strains that were sequenced previously. Initial characterization in genomic relatedness based on ANI distances of the chromosome

showed that the total set of strains were grouped in three main clades: A (397 strains), B (12 strains) and C (3 strains), with clade A containing the majority of strains, similar to what has been previously described with other typing methods such as multiple-locus variable-number tandem repeat analysis (MLVA) (Figure 3.2). When we compared the clustering profile for both plasmids versus that of the chromosome, we observed a high topological correlation. To quantify the strength of the correlation we used two metrics: the cophenetic distance, defined as the intergroup distance at which two observations are first combined into a single cluster, and the Baker's Ω index, defined as the rank correlation between the stages at which pairs of observations combine in each one of the two dendrograms being compared. For pXO1, the calculated cophenetic correlation was 0.70 and the Baker's Ω index correlation was 0.62. For pXO2, the calculated cophenetic correlation was 0.89 and the Baker's Ω correlation was 0.93, which indicated that, in general, pXO1 and pXO2 phylogenies recapitulate that of the chromosome.

3.4.5 Gene content variation of pXO1 and pXO2

To avoid limitations of the assembly process such as gaps or truncated genes and mis-assemblies, we assessed gene content variation of the plasmids by recruiting high-quality (trimmed) Illumina reads against the predicted genes on the plasmid and determining gene presence/absence by the number of reads recruited (or not) on the gene. Genomes containing one or both plasmids showed a highly conserved gene content in general (Figure 3.4AB).

The calculated pXO1 pangenome was composed of 197 orthologous genes; 179 of them (91%) were present in all strains (strict core), 195 (99%) were present in at least 95% of all the strains (relaxed core) and only two genes were found to be variable in pXO1. 108 genes composed the pXO2 pangenome; 96 genes were part of the strict core (89%) while 102 genes were part of the relaxed core. Only 6 genes composed the variable genome. Although no large plasmid gene content diversity was generally observed between any two genomes analyzed, we identified a large fragment deletion in the pXO1 plasmid of one strain: 2000031682 (Figure 3.4C). The deleted fragment was about ~46.3 Kb in length and contained 39 genes in total, including the main virulence factors responsible for anthrax toxin: *cya*, *pagA* and *lef* and the transcriptional activator

atxA. We also identified a number of integrases, resolvases and transposases in and around the deleted fragment (Figure 3.4D). Re-sequencing and re-assembling of this particular strain confirmed the large gene deletion.

In addition, we identified orthologous genes present in both plasmids that showed higher sequencing depth compared to the majority of plasmid genes. These genes were most likely multi-copy genes. In pXO1 we identified three multi-copy genes that were observed to have two copies, on average, consistently across the complete set of strains. In pXO2, we identified three genes with consistent multi-copy pattern in most of the strains characterized in this analysis. In both plasmids, these genes corresponded to transposases for insertion sequence elements (IS).

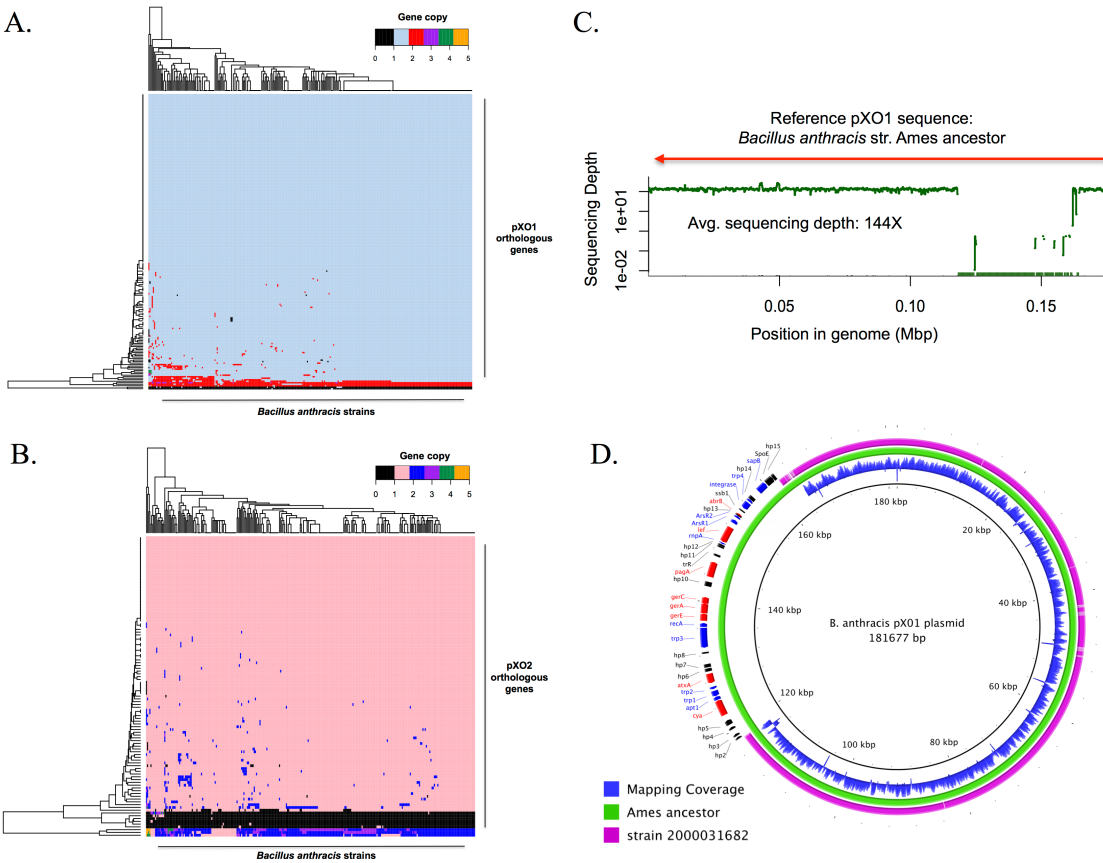


Figure 3.4. Gene content variation of pXO1 and pXO2. (A) Hierarchical clustering of *B. anthracis* strains containing plasmid pXO1 (columns) based on the estimated sequencing depth for each representative orthologous gene (rows) normalized by the median sequencing depth for each genomic library (columns). (B) Shows the same as (A) but for plasmid pXO2. (C) Read recruitment plot showing the absence of read coverage in strain 2000031682 in a region of ~ 46.3Kbp while the calculated average sequencing depth of the covered region was 144X. (D) Circular plot comparing pXO1 in *B. anthracis* strain Ames ancestor (in green) and strain 2000031682 (in purple). Mapping coverage from sequencing reads of strain 2000031682 along the plasmid is shown in blue (innermost circle). The deleted fragment is shown along with the functional annotation of the genes identified in the region. Red arrows denote the position and strand of anthrax virulence determinant genes identified in the missing region while blue arrows identify mobile elements and black arrows denote genes encoding hypothetical proteins.

3.4.6 Comparison to pXO1/pXO2-like plasmids of other members of the *B. cereus sensu lato* group

To increase understanding of the evolutionary relationships of *B. anthracis* plasmids with those of other (non-*B. anthracis*) members of the *B. cereus sensu lato* group, we performed an ANI-based clustering analysis of selected strains from our dataset together with available reference *B. cereus s.l.* strains. The final set included: 94 *B. cereus* strains, 11 representative *B. anthracis* strains identified from our dataset, 11 *B. thuringiensis* strains and one *B. mycoides* isolate. Results showed that strains clustered in 3 main groups: clade I, clade II and clade III (Figure 3.5A). Representative strains from *B. anthracis* were grouped in clade I, the same group where several *B. cereus* isolates of clinical origin were placed. The majority of *B. thuringiensis* strains (7 out 11) were grouped in cluster III, although four of them were assigned to clade I. These relationships were consistent with previous phylogeny characterization based on MLST schemes or chromosomal core proteins, which have shown that *B. cereus*, *B. mycoides* and *B. thuringiensis* are not confined within discrete clades and are therefore, not monophyletic species (26-28).

We then attempted to identify non-*B. anthracis* genomes that carried a complete or partial genomic backbone with pXO1 and/or pXO2. To achieve this goal, we followed two approaches since our genome sequences were incomplete (draft). First, we identified large contigs (≥ 500bp) with ≥ 80 % identity and ≥ 80 % sequence coverage to reference pXO1 and pXO2 plasmids from *B. anthracis* Ames Ancestor (we called these contigs 'pXO1/2-like contigs'); and second, we generated read recruitment plots to visualize and quantify the sequencing depth coverage provided by reads of the genomic library of the corresponding strain along the reference plasmid sequence (see Methods and Materials). We identified 33 genomes containing pXO1-like contigs, 12 of which were assignable to clade I, two to clade II, and 20 to clade III. We also identified 17 strains containing pXO2-like contigs; four were assignable to clade I, three to clade II, and 10 to clade III (e.g., Fig. 3.5A). Functional characterization of the genes predicted in pXO1-like contigs (365 in total) showed that the majority of those genes were hypothetical proteins (97%) and only 42 genes (3%) could be functionally annotated. From these, five genes encoded for hyaluronan synthase (Figure 3.5B). However, no anthrax toxin genes were identified among these sequences.

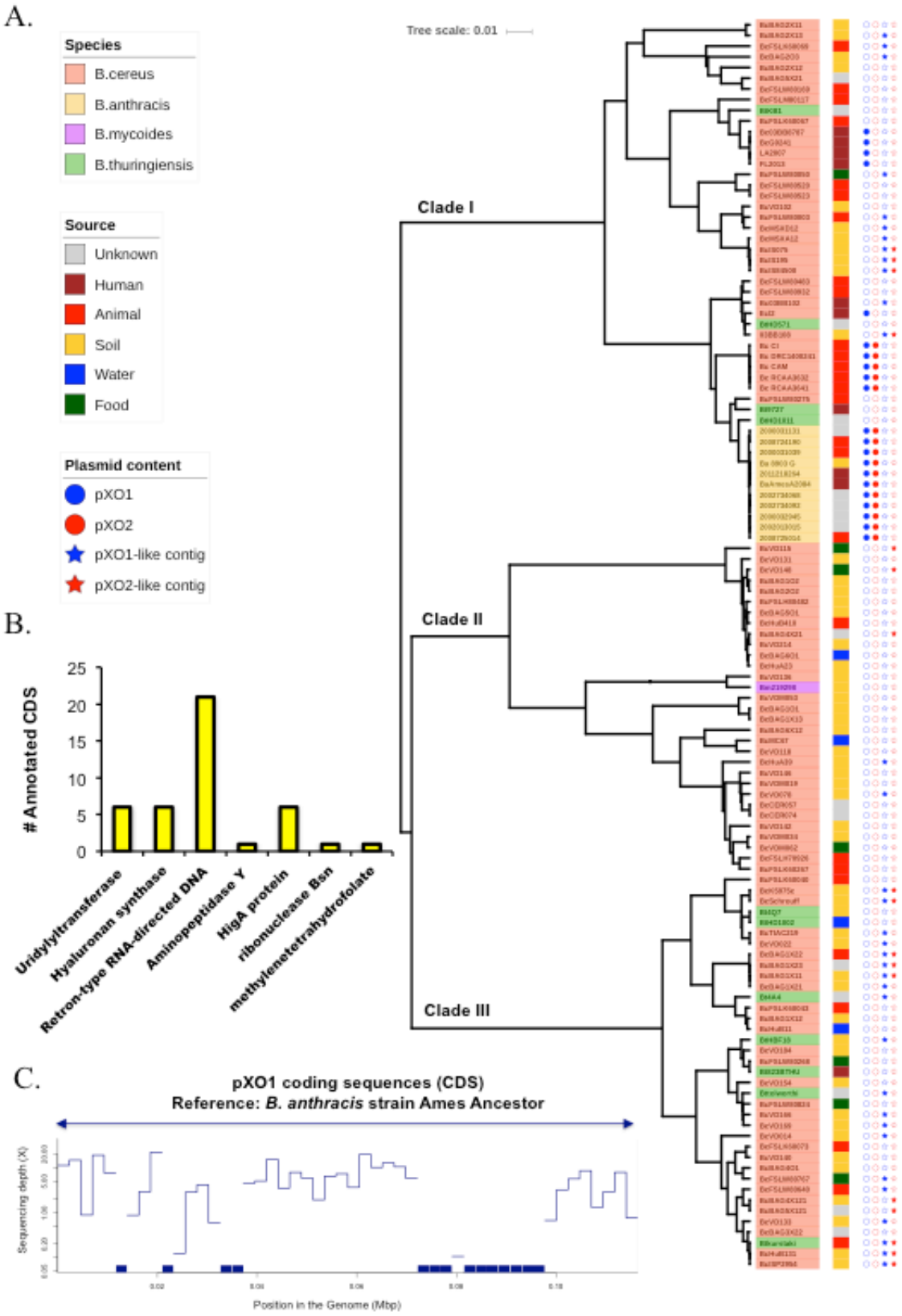
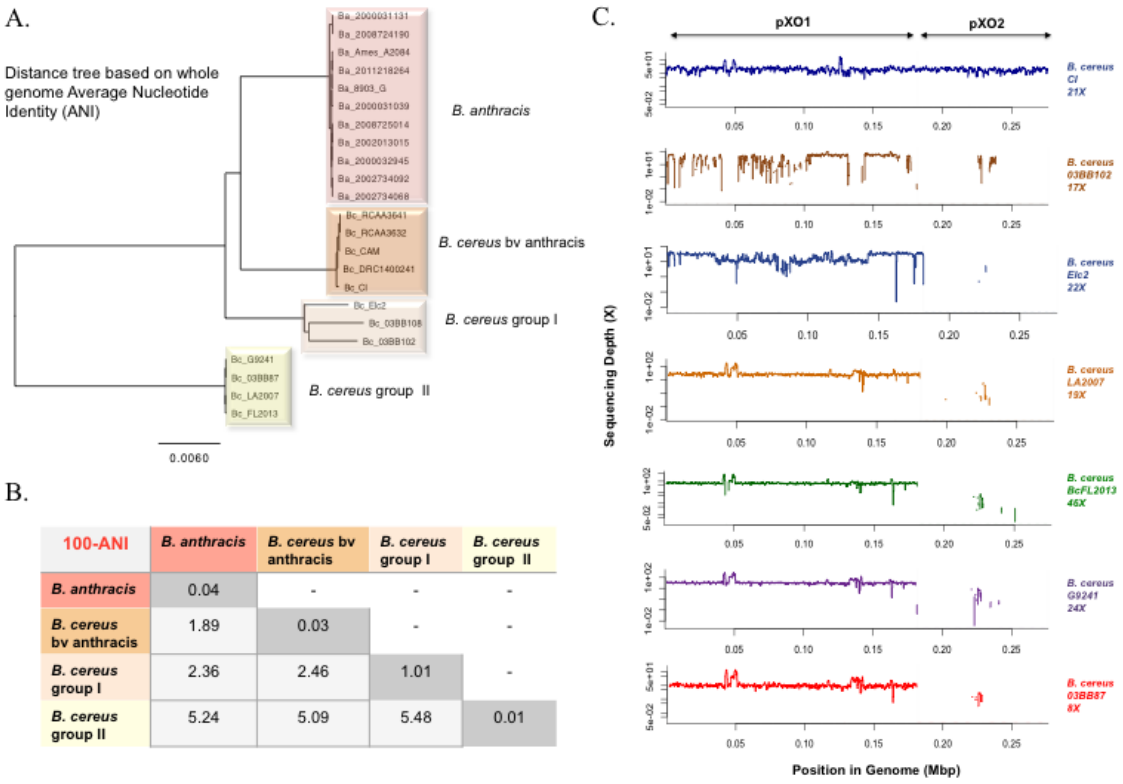

A.
Species
B.cereus
B.anthracis
B.mycoides
B.thuringiensis
Source
Unknown
Human
Animal
Soil
Water
Food
Plasmid content
pXO1
pXO2
pXO1-like contig
pXO2-like contig
Tree scale: 0.01
Clade I
Clade II
Clade III
B.
Annotated CDS
25
20
15
10
5
0
Uridylyltransferase
Hyaluronan synthase
Retron-type RNA-directed DNA
Aminopeptidase Y
HigA protein
ribonuclease Bsn
methylenetetrahydrofolate
C.
pXO1 coding sequences (CDS)
Reference: B. anthracis strain Ames Ancestor
Sequencing depth (X)
Position in the Genome (Mbp)

Figure 3.5. Genomic characterization of *Bacillus cereus sensu lato* group. (A) Clustering of species within *Bacillus cereus sensu lato* group based on ANI distances. *B. anthracis* members are colored in yellow; *B. cereus sensu stricto* strains are colored in pink; *B. thuringiensis* members are colored in green and the single representative *B. mycoides* is colored in purple. The right colored strip indicates the source of isolation. Filled or empty circles indicate the presence (filled) or absence (empty) of pXO1 (in blue) and pXO2 (in red). Filled stars denote the strains for which pXO1-like (blue) or pXO2-like (red) contigs were identified. (B) The number of functionally annotated (excluding hypothetical) coding sequences predicted in pXO1-like contigs (42 in total, 3% of the total number). (C) Example of *B. cereus* strain VD014 showing partial pXO1-like backbone homologous to pXO1 from *B. anthracis* Ames Ancestor.

We also added to the analysis described above 11 previously characterized pathogenic *B. cereus* genomes carrying complete pXO1 and/or pXO2 plasmids. In particular, five *B. cereus* biovar anthracis strains (RCA_A_364-1, RCA_A_363-2, DRC_14-0024-1, CAM and CI), which were isolated from dead mammals with an illness consistent with anthrax (chimpanzees, gorillas, elephants and goats) in West and Central Africa (23), and six pathogenic *B. cereus* strains isolated from human cases of pneumonia or cutaneous lesions (G9241, BcFL2013, LA2007, 03BB87, 03BB102 and Elc2). These strains were compared to 11 representative *B. anthracis* genomes (from our dataset) and an additional *B. cereus* strain (03BB108, isolated from dust at the same worksite where a Texas welder contracted fatal pneumonia in 2003), that carried partial homology to the backbone of pXO1.

Clustering analysis based on ANI dissimilarity revealed four groups (Figure 3.6A): one highly clonal group composed by *B. anthracis* strains with 0.04 average intra-group ANI distance (i.e., 99.96 identity), the second group was composed by *B. cereus* biovar anthracis isolates which was also highly similar with an average of 0.03 ANI distance. The third group, which we labeled as *B. cereus* group I, was composed by three *B. cereus* isolates (03BB108, 03BB102 and Ecl2) and had 1.01 average ANI distance among them. Finally, the fourth group, labeled as *B. cereus* group II, was composed of four human-pathogenic *B. cereus* strains (LA2007, BcFL2013, G9241 and 03BB87) and showed an average 0.01 intragroup ANI distance, the smallest observed

intragroup diversity. *B. cereus* group II was the most divergent to all other groups, with an average inter-group ANI distance of 5.27 (Figure 3.6B).



100-ANI	*B. anthracis*	*B. cereus* bv anthracis	*B. cereus* group I	*B. cereus* group II
B. anthracis	0.04	-	-	-
B. cereus bv anthracis	1.89	0.03	-	-
B. cereus group I	2.36	2.46	1.01	-
B. cereus group II	5.24	5.09	5.48	0.01

Figure 3.6. Genomic characterization of *B. cereus* strains carrying complete anthrax-like plasmids. (A) Whole genome ANI distance tree between representative *B. anthracis* strains and pathogenic *B. cereus* strains carrying anthrax-like plasmids. (B) Average intra- and inter-group ANI distance calculated for the set of strains carrying anthrax or anthrax-like plasmids. (C) Read-based detection and characterization of plasmids pXO1 and pXO2 in anthrax-like *B. cereus* strains isolated from human and mammal tissues in the West and Central Africa and North America. Strain CI was chosen as a representative strain for the *B. cereus biovar anthracis* group.

Plasmid detection and quantification based on read coverage confirmed that all *B. cereus* biovar anthracis strains carried complete pXO1 and pXO2-like plasmids while

strains BcFL2013, G9241, 03BB87, LA2007 and Elc2 harbored a complete pXO1-like but not a pXO2-like plasmid (Figure 3.5C). 03BB102, which was isolated from a fatal pneumonia in TX, differed from these other strains in that it does not harbor a full length pXO1 or pXO2-like plasmid, although partial sequence homology to the pathogenicity island was detected (51.72% of the total genes of the island were present). (22). Our sequence-based analysis revealed that 03BB102 harbors the typical anthrax virulence genes but lacks about half of the canonical pXO1 gene content (Figure 3.5C).

Further gene-based characterization showed that this plasmid (p03BB102) carried a complete, 5.1 Kbp pXO2 *cap* locus with 93% nucleotide identity to the Ames strain *cap* locus (accession number AE017335) and was flanked by IS elements. In addition, duplicate homologs of protective antigen genes (*pagA* and *pagR*) were also identified; one homolog showed ~99% nucleotide identity to its ortholog in pXO1 Ames Ancestor while the second showed 92%-94% (respectively), indicating that these homologs have already begun to diverge (Figure 3.7). Although pXO1/pXO2-like plasmids seem to be remarkably conserved in terms of gene content and synteny, strain 03BB102 is an exception to this rule, which suggests that plasmid diversity in nature may be higher than previously thought.

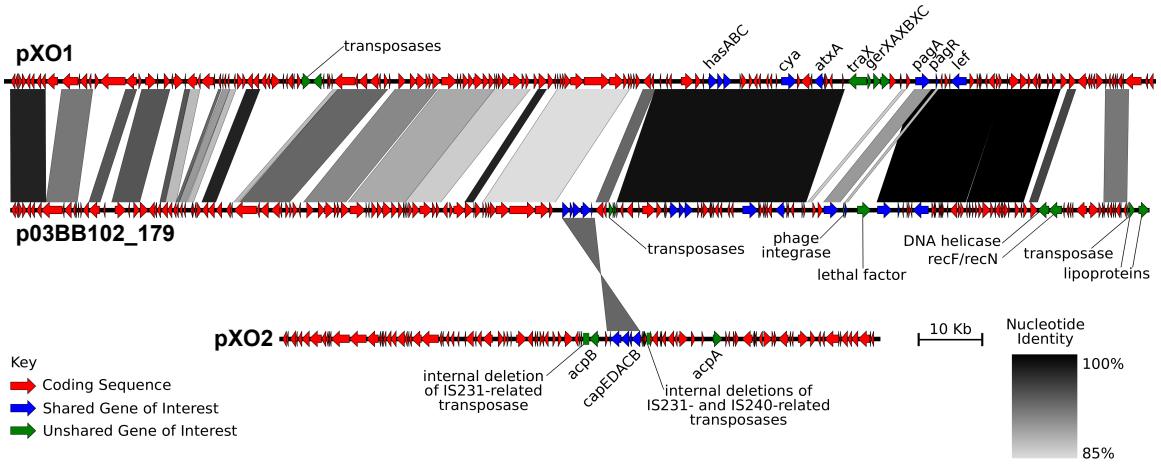


Figure 3.7. Gene content comparison between *B. anthracis* pXO1/pXO2 and *B. cereus* strain 03BB102 plasmids. The connecting lines show the presence and location of shared genes while the gray scale represents the level of nucleotide identity. p03BB102 carries a *cap* locus (~5.1 Kbp) that is 93% nucleotide identity to the *cap* locus

in reference pXO2 (*B. anthracis* 'Ames Ancestor' accession number AE017335) and is flanked by IS elements. A duplicate homolog of *pagA* and *pagR* genes is also present in p03BB102; one homolog showing 99% identity compared to its pXO1 *B. anthracis* homolog (*B. anthracis* 'Ames Ancestor' accession number AE017336) for both genes, while the second shows 92% and 94% respectively.

In addition, we also calculated plasmid copy number in the set of *B. cereus* strains carrying anthrax-like plasmids (Table 3.1). Sequencing breath (or breadth of coverage) was calculated as the percentage of bases of the reference *B. anthracis* Ames Ancestor plasmid sequences that were covered by reads of the corresponding genome (rows) at ≥ 2X sequencing depth. Plasmid copy number was estimated as described in the Materials and Methods section. The estimated average pXO1 copy number in *B. cereus* biovar anthracis strains was 1.8 while for the set of human-pathogenic *B. cereus* was 2.32, which was similar to the estimated average for *B. anthracis* (3.86). For pXO2, the estimated average copy number in *B. cereus* biovar anthracis was 2.12, similar to the estimated copy number in *B. anthracis* (2.29).

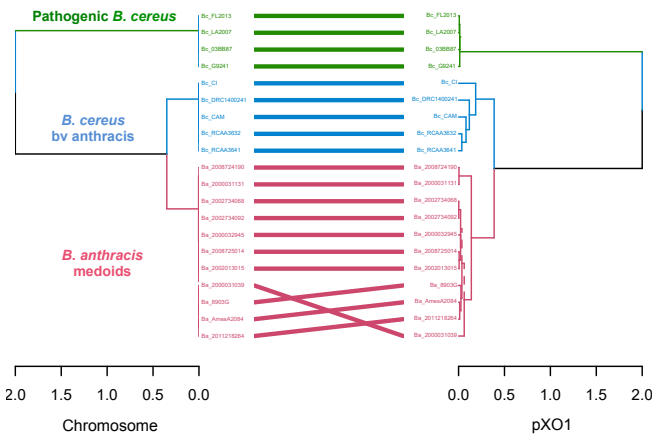
Table 3.1. Sequencing breadth and copy number variation of pXO1 and pXO2 plasmids in five *B. cereus* strains biovar anthracis and five human-pathogenic *B. cereus*.

Strain ID	*Group*	*pXO1 Seq. breadth (%)*	*pXO2 Seq. breadth (%)*	*pXO1 copy number*	*pXO2 copy number*
CI	*B. cereus bv. anthracis*	99.90	99.35	1.3	1.1
CAM	*B. cereus bv. anthracis*	99.91	99.17	1.8	0.8
DRC_14-0024-1	*B. cereus bv. anthracis*	77.92	98.95	1.3	5.6
RCA_A_364-1	*B. cereus bv. anthracis*	99.94	99.48	2.3	1.6
RCA_A_363-2	*B. cereus bv. anthracis*	99.94	99.48	2.3	1.5
FL2013	*B. cereus G-II*	96.71	1.7	4.4	-
G9241	*B. cereus G-II*	97.90	1.9	1.4	-
03BB87	*B. cereus G-II*	96.89	1.8	1.0	-
LA2007	*B. cereus G-II*	97.98	1.8	2.5	-
03BB102	*B. cereus G-I*	51.72	6.5	-	-

3.4.7 Assessing origin and vertical vs. horizontal transmission of plasmids

To determine if the pXO1/pXO2-like plasmids have been transferred between members of *B. cereus* and *B. anthracis*, we contrasted the phylogenetic relationships among the genomes based on the chromosomal genes relative to those of the plasmids. Phylogenetic reconstruction based on plasmid core orthologous genes of the strains harboring a complete pXO1 (139 genes) and/or pXO2 (88 genes) showed a similar topology to the one observed with whole genome ANI distance based on tanglegram analysis (Figure 3.8A, 3.9B), indicating limited mobilization of the plasmids among the genome. For pXO1, *B. cereus* biovar anthracis strains were closer to the *B. anthracis* group than they were to the set of human-pathogenic *B. cereus* group II, and Elc2 was the most divergent strain. For pXO2, three main clades were observed, one containing all *B. anthracis* strains, one containing strain *B. cereus* CI and one clade containing the remaining *B. cereus* biovar anthracis strains (RCA_A_364-1, RCA_A_363-2, DRC_14-0024-1, and CAM). Strain CI was more similar to the *B. anthracis* group than it was to the other *B. cereus* biovar anthracis isolates.

A.



B.

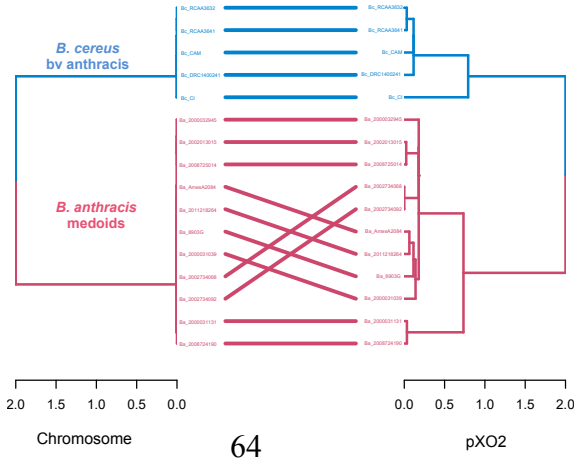


Figure 3.8. Assessment of plasmid lateral transfer between representative *B. anthracis* and pathogenic *B. cereus* strains carrying complete pXO1/pXO2-like plasmids. Comparison of phylogenetic relationships based on core genome for the chromosome and pXO1 (A) and for the chromosome and pXO2 (B) in strains carrying one or both plasmids. Phylogenetic reconstructions in panel A were based on the alignment of 210,123 variable positions found in the concatenated alignment of 4,233 core orthologous genes for the chromosome and 458 variable positions identified in the concatenated alignment of 149 pXO1 core orthologous genes. Phylogenetic relationships in panel B were constructed from the alignment of 74,389 variable positions found in the concatenated alignment of 4,616 core orthologous genes for the chromosome and 120 variable positions identified in the concatenated alignment of 88 pXO2 core orthologous genes. No signal of plasmid lateral transfer between both phylogroups was apparent.

Further, clustering based on the presence/absence of the variable genes of both plasmids showed a similar grouping pattern to that of the chromosome, indicating that (higher) gene-content variation largely correlates to (higher) genome divergence. Collectively, these results indicated limited horizontal transfer of the plasmid between *B. cereus* and *B. anthracis*. Accordingly, *B. cereus* genomes harboring the *B. anthracis* plasmids appear to have maintained these plasmids since their last common ancestor with *B. anthracis*. However, we did observe topological incongruences between the chromosome and pXO1 core gene trees within *B. anthracis* (entanglement = 0.20), indicating that the plasmid might have undergone horizontal transfer within the group. For instance, the chromosome- versus plasmid-based topologies were significantly incongruent by all three tests applied, i.e., the one-sided Maximum Likelihood Kishino-Hasegawa test (KH) (57), the Shimodaira-Hasegawa test (SH) (56) and the expected likelihood weight test (ELW) (58) (pval-1sKH = 0.005, pval-SH = 0.002, c-ELW = 0.002; all tests were applied with 5% significance level and 1,000 re-samplings using the RELL method). However, individual gene-level assessment showed that the abovementioned topological inconsistence was predominantly due to recombination and/or varied selection pressures in only five genes, and was not plasmid-wide. When these genes

were removed from the core gene alignment, the plasmid tree and the chromosome were topologically more congruent (entanglement = 0.05). We also observed that the complete pXO1 and chromosome trees grouped in the same cluster (less distance between them) in a low-dimensional representation of the topological variability of the trees evaluated using the Kendall and Colijn metric (see methods and materials). Collectively, our analyses revealed no strong evidence of plasmid lateral transfer between or within *B. anthracis* and *B. cereus*.

3.5 Discussion

In this study, we estimated plasmid copy number in a large collection of newly sequenced *B. anthracis* strains, characterized their full plasmid gene content, and compared the phylogenetic diversity of representative genomes with other *Bacillus* species carrying complete or partial pXO1/pXO2-like plasmids. Our major findings revealed that *B. anthracis* cells maintain, on average, 3.86 copies of pXO1 and 2.29 copies of pXO2, and reveal that there is positive linear correlation in the number of copies between both plasmids, which was consistent with two previously reported sequence-based studies (12, 13). The gene content of these anthrax-plasmids is remarkably stable although a few genomes (e.g., strain 2000031682) lacked large parts of the plasmids. Further, the number of plasmid copies that *B. anthracis* genomes harbor seems influenced by the source from where the strains were isolated (animal or environmental) but not phylogeny. We also identified several environmental *B. cereus s.l.* strains containing pXO1 and pXO2-like plasmids, some previously reported (31). We found no strong evidence of plasmid exchange between *B. anthracis* and *B. cereus s.l.* genomes, which suggests plasmid maintenance since the last common ancestor of the two species. We noted, however, statistically significant topological incongruences between the chromosome tree and those of a few individual plasmid genes within *B. anthracis*, indicating that horizontal exchange and/or intra-genomic recombination of individual plasmid genes has occurred in the recent evolutionary time.

Our estimates reveal a lower number of pXO1 copies per chromosome, on average (n = 3.86), compared with earlier studies based on molecular methods such as qPCR. For example, Coker *et al.* (2003) estimated ratios of up to 40 copies of pXO1

while Pilo *et al.* (2011b) reported 10-11 copies of the same plasmid. In both cases, the estimation based on quantities of a portion of a single gene per replicon, representing ~0.1% of the total replicon length, seem to be inflated compared with our more comprehensive shotgun sequence-based estimations. However, for pXO2, both qPCR and sequencing provided similar estimates of approximately 1-2 copies per cell, indicating that PCR may have overestimated pXO1 abundance (since the competing hypothesis that sequencing was biased against pXO1 abundance but not that of pXO2 appears to be less parsimonious), although we have also identified strains carrying up to 4.5 copies of pXO2. The fact that plasmid copy number is not a phylogenetically conserved trait indicates that extrinsic forces, perhaps environmental factors such as temperature, pH, soil moisture, cation levels, among others, might play a more important role in determining the number of replicons that *B. anthracis* cells maintain. In other words, the plasmid copy number is a trait that may be adjusted in response to environmental cues. Studies of the ecology of *B. anthracis* have shown that the global distribution of anthrax is largely determined by climatic factors and land features, where, for example, soils with high calcium levels and a pH above 6.1 foster better spore survival (6, 9, 29).

In this study the gene content diversity of the plasmids across a set of strains carrying one or both plasmids was characterized. Our results confirmed the highly conserved gene content and synteny for both plasmids (>97% of total plasmid genes shared) similar to what has been previously described for this species. In addition, we identified a single strain (2000031682) with a large fragment deletion in the pXO1 plasmid. The deleted fragment size was approximately 46.3 Kb and carried the main virulence genes responsible for anthrax toxin production: *cya*, *pagA, lef* and *atxA*. While the history of this strain is not clear, it was originally archived at CDC in 1964 on an agar slant and stored at room temperature. The strain was recovered from the slant and frozen at -70^{o}C in 2001. We previously reported that numerous strains in this collection were cured of plasmids during decades of room temperature storage (54). We cannot ascertain whether this stain was received in this condition or if the deletion may have occurred during storage.

Environmental *B. cereus s.l.* strains possessing partial or complete pXO1/pXO2-like plasmids, other than *B. cereus* biovar anthracis, which encoded complete *B. anthracis* plasmids were also studied. Through bioinformatic approaches, we identified

50 strains having contigs homologous to pXO1 and/or pXO2. We confirmed that pXO1 and pXO2-like contigs are widely prevalent in environmental isolates of the *B. cereus s.l.* group, similar to what was previously revealed by Van der Auwera *et al*. (2013) using PCR-based approaches (31). The annotation of the genes present in pXO1/pXO2-like contigs showed that most of them were hypothetical proteins, with few of them predicted to be involved in DNA insertion and transposition (for example retron-type RNA-directed DNAases and ribonucleases; see Fig 4B) but no genes encoding for anthrax toxins were identified. However, we found genes encoding for hyaluronic acid (HA) capsule formation in these contigs, a trait that provides pathogenic bacilli with capsular material to escape the innate host immune response and is involved in the pathogenesis of anthrax-like diseases. Given that all these *B. cereus s.l.* strains analyzed here were of environmental origin, these findings might indicate that at least some of the virulence factors encoded on the *B. anthracis* plasmids (e.g., HA capsule formation) may be important for survival in the environment outside the human host.

A potential limitation in our study is that some strains could have lost some or all of their plasmids during successive subculturings. The rate of plasmid loss during recurrent subculturing could have been accelerated under stressing laboratory conditions (54), which might potentially have biased our estimation of the 'true' copy number variation. To minimize this issue, the original culture stock was used for preparing DNA for sequencing, not derived subcultures. Further, a larger effort on soil field sampling would be necessary to evaluate how frequently *B. cereus* group strains carry complete or partial pXO1/pXO2-like plasmids and if these are exchanged or not with *B. anthracis*.

3.6 Conclusions and recommendations

In summary, by using next generation sequencing data, we have estimated *B. anthracis* plasmid copy number, characterized their genomic diversity and compared representative strains with clinical and environmental *B. cereus s.l.* strains carrying pXO1/pXO2-like plasmids. The results derived from this study advance our understanding of the biology of the *B. cereus* group, improve the ecological and evolutionary framework used to classify species, and appropriately define phylogenetic

relationships and taxonomic assignments within the *B. cereus s.l.* group. Our results also highlighted the advantages of using genomic relatedness (as measured by ANI, for example), instead of plasmid-encoded traits, to assign taxonomy and robustly resolve the relationships among closely related members of the *B. cereus s.l.* group. These results and interpretations were also consistent with previous studies of plasmid-encoded traits in other bacterial species such as *Clostridium botulinum* [e.g.,59]. Therefore, the results derived from our study will help to improve the ecological and evolutionary framework used to classify species and appropriately define phylogenetic relationships, especially in a bacterial group that exhibits high phenotypic diversity as *Bacillus cereus s.l*.

Although the collection size of the *B. anthracis* strains sequenced in this study and the number of strains deposited in public databases is likely still small compared to the total natural diversity of the species, to the best of our knowledge, this is the largest study characterizing *B. anthracis* plasmid copy number variation and their gene content diversity using sequencing data to date. Hence, the data presented here should facilitate future studies of *B. anthracis* and its virulence plasmids. Finally, the bioinformatic approaches used in this study can also be applied as a reference framework for epidemiological studies involving this and other microorganisms of medical relevance.

3.7 Acknowledgements

This work was supported by US National Science Foundation award No 1356288 and DHHS/PHS/CDC award No RF023 to KTK. A.P.G. was partially supported by Colciencias -Colombian Administrative Department for Science, Technology and Innovation through a doctoral fellowship. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. Mention of company names or products does not constitute endorsement by the CDC.

3.8 References

1. Derzelle S, Girault G, Kokotovic B, Angen Ø. 2015. Whole Genome-Sequencing and Phylogenetic Analysis of a Historical Collection of Bacillus anthracis Strains from Danish Cattle. PloS one 10:e0134699.
2. Van Ert MN, Easterday WR, Huynh LY, Okinaka RT, Hugh-Jones ME, Ravel J, Zanecki SR, Pearson T, Simonson TS, U'Ren JM. 2007. Global genetic population structure of Bacillus anthracis. PloS one 2:e461
3. Girault, G, Blouin, Y, Vergnaud, G, Derzelle, S. 2014. High-throughput sequencing of Bacillus anthracis in France: investigating genome diversity and population structure using whole-genome SNP discovery. BMC genomics 15: 288-294.
4. Hendricks KA, Wright ME, Shadomy SV, Bradley JS, Morrow MG, Pavia AT, Rubinstein E, Holty JE, Messonnier NE, Smith TL. 2014. Centers for Disease Control and Prevention expert panel meetings on prevention and treatment of anthrax in adults. Emerg Infect Dis 20:e130687.
5. Pilo P, Frey J. 2011a. Bacillus anthracis: Molecular taxonomy, population genetics, phylogeny and patho-evolution. Infection, Genetics and Evolution 11:1218-1224.
6. Hugh-Jones M, Blackburn J. 2009. The ecology of Bacillus anthracis. Molecular aspects of medicine 30:356-367
7. Riojas MA, Kiss K, McKee ML, Hazbón MH. 2015. Multiplex PCR for Species-Level Identification of Bacillus anthracis and Detection of pXO1, pXO2, and Related Plasmids. Health security 13:122-129.
8. Coker PR, Smith KL, Fellows PF, Rybachuck G, Kousoulas KG, Hugh-Jones ME. 2003. Bacillus anthracis virulence in Guinea pigs vaccinated with anthrax vaccine adsorbed is linked to plasmid quantities and clonality. Journal of clinical microbiology 41:1212-1218.
9. Bergman NH, editor. 2011. Bacillus anthracis and Anthrax. John Wiley & Sons.
10. Irenge LM, Gala J-L. 2012. Rapid detection methods for Bacillus anthracis in environmental samples: a review. Applied microbiology and biotechnology 93:1411-1422.
11. Pilo P, Rossano A, Bamamga H, Abdoulkadiri S, Perreten V, Frey J. 2011b. Bovine Bacillus anthracis in Cameroon. Applied and environmental microbiology 77:5818-5821.
12. Straub T, Baird C, Bartholomew RA, Colburn H, Seiner D, Victry K, Zhang L, Bruckner-Lea CJ. 2013. Estimated copy number of Bacillus anthracis plasmids pXO1 and pXO2 using digital PCR. Journal of microbiological methods 92:9-10.
13. Ravel J, Jiang L, Stanley ST, Wilson MR, Decker RS, Read TD, Worsham P, Keim PS, Salzberg SL, Fraser-Liggett CM. 2009. The complete genome sequence of Bacillus anthracis Ames “Ancestor”. Journal of bacteriology 191:445-446.
14. Papazisi L, Rasko DA, Ratnayake S, Bock GR, Remortel BG, Appalla L, Liu J, Dracheva T, Braisted JC, Shallom S. 2011. Investigating the genome diversity of B. cereus and evolutionary aspects of B. anthracis emergence. Genomics 98:26-39.
15. Okinaka RT, Keim P. 2016. The phylogeny of Bacillus cereus sensu lato. Microbiology spectrum 4
16. Økstad OA, Kolstø A-B. 2011. Genomics of bacillus species, p 29-53, Genomics of foodborne bacterial pathogens. Springer.

17. Granum PE, Lund T. 1997. Bacillus cereus and its food poisoning toxins. FEMS microbiology letters 157:223-228.
18. Helgason E, Caugant DA, Olsen I, Kolstø A-B. 2000. Genetic structure of population of Bacillus cereus and B. thuringiensis isolates associated with periodontitis and other human infections. Journal of clinical microbiology 38:1615-1622.
19. Maughan H, Van der Auwera G. 2011. Bacillus taxonomy in the genomic era finds phenotypes to be essential though often misleading. Infection, Genetics and Evolution 11:789-797
20. Marston CK, Ibrahim H, Lee P, Churchwell G, Gumke M, Stanek D, Gee JE, Boyer AE, Gallegos-Candela M, Barr JR. 2016. Anthrax toxin-expressing *Bacillus cereus* isolated from an anthrax-like eschar. PloS one 11:e0156987
21. Hoffmaster AR, Ravel J, Rasko DA, Chapman GD, Chute MD, Marston CK, De BK, Sacchi CT, Fitzgerald C, Mayer LW. 2004. Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. Proceedings of the National Academy of Sciences of the United States of America 101:8449-8454.
22. Hoffmaster AR, Hill KK, Gee JE, Marston CK, De BK, Popovic T, Sue D, Wilkins PP, Avashia SB, Drumgoole R. 2006. Characterization of *Bacillus cereus* isolates associated with fatal pneumonias: strains are closely related to Bacillus anthracis and harbor B. anthracis virulence genes. Journal of clinical microbiology 44:3352-3360.
23. Antonation KS, Grützmacher K, Dupke S, Mabon P, Zimmermann F, Lankester F, Peller T, Feistner A, Todd A, Herbinger I. 2016. Bacillus cereus Biovar Anthracis Causing Anthrax in Sub-Saharan Africa—Chromosomal Monophyly and Broad Geographic Distribution. PLoS Negl Trop Dis 10:e0004923.
24. Klee SR, Özel M, Appel B, Boesch C, Ellerbrok H, Jacob D, Holland G, Leendertz FH, Pauli G, Grunow R. 2006. Characterization of Bacillus anthracis-like bacteria isolated from wild great apes from Cote d'Ivoire and Cameroon. Journal of bacteriology 188:5333-5344.
25. Pena-Gonzalez A, Marston CK, Rodriguez-R LM, Kolton CB, Garcia-Diaz J, Theppote A, Frace M, Konstantinidis KT, Hoffmaster AR. 2017. Draft genome sequence of *Bacillus cereus* LA2007, a human-pathogenic isolate harboring anthrax-like plasmids. Genome Announc 5:e00181-17. https://doi.org/10.1128/genomeA.00181-17.
26. Zwick ME, Joseph SJ, Didelot X, Chen PE, Bishop-Lilly KA, Stewart AC, Willner K, Nolan N, Lentz S, Thomason MK, Sozhamannan S, et al. 2012. Genomic characterization of the Bacillus cereus sensu lato species: backdrop to the evolution of Bacillus anthracis. Genome research, 22:1512-1524.
27. Tourasse NJ, Kolstø AB. 2008. SuperCAT: a supertree database for combined and integrative multilocus sequence typing analysis of the Bacillus cereus group of bacteria (including B. cereus, B. anthracis and B. thuringiensis). Nucleic acids research 36:D461-468.
28. Tourasse NJ., Helgason E, Økstad OA, Hegna IK , KolstØ AB. 2006. The Bacillus cereus group: novel aspects of population structure and genome dynamics. Journal of applied microbiology, 101:579-593.
29. World Health Organization and International Office of Epizootics. 2008. *Anthrax in humans and animals*. World Health Organization.
30. Okinaka RT, Price EP, Wolken SR, Gruendike JM, Chung WK, Pearson T, Xie G, Munk C, Hill KK, Challacombe J, Ivins BE. 2011. An attenuated strain of Bacillus

anthracis (CDC 684) has a large chromosomal inversion and altered growth kinetics. BMC Genomics 12:477-479.
31. Van der Auwera GA, Feldgarden M, Kolter R, Mahillon J. 2013. Whole-genome sequences of 94 environmental isolates of Bacillus cereus sensu lato. Genome announcements, 1:pp.e00380-13.
32. Buffalo V. 2014. Scythe–A Bayesian Adapter Trimmer (Version 0.994 BETA).
33. Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC bioinformatics 11:485
34. Peng Y, Leung HCM, Yiu S-M, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420-1428
35. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research 25:1043-1055
36. Rodriguez-R LM, Konstantinidis KT. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints.
37. Lukashin AV, Borodovsky M. 1998. GeneMark.hmm: new solutions for gene finding. Nucleic acids research 26:1107-1115.
38. Luo C, Rodriguez-r LM, Konstantinidis KT. 2014. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. Nucleic acids research:gku169.
39. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. International journal of systematic and evolutionary microbiology. 57:81-91.
40. Konstantinidis KT, Tiedje JM. 2005. Towards a genome-based taxonomy for prokaryotes. Journal of bacteriology 187:6258-6264.
41. Kaufman L, Rousseeuw PJ. 1990. Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis 68-125.
42. Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20:53-65.
43. Murtagh F, Legendre P. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?. Journal of Classification 31:274-295.
44. Blomberg SP, Garland Jr T, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717-745.
45. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463-1464.
46. Galili T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics:btv428.
47. Baker FB. 1974. Stability of two hierarchical grouping techniques Case I: Sensitivity to data errors. Journal of the American Statistical Association 69:440-445.
48. Sokal RR, Rohlf FJ. 1962. The comparison of dendrograms by objective methods. Taxon 1:33-40.
49. Weigand MR, Pena-Gonzalez A, Shirey TB, Broeker RG, Ishaq MK, Konstantinidis KT, Raphael BH. 2015. Implications of genome-based

discrimination between Clostridium botulinum group I and Clostridium sporogenes strains for bacterial taxonomy. Applied and environmental microbiology 81:5420-5429.
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of molecular biology 215:403-410.
51. Beckett S, Jee J, Ncube T, Pompilus S, Washington Q, Singh A, Pal N. 2014. Zero-inflated Poisson (ZIP) distribution: parameter estimation and applications to model data from natural calamities. Involve, a Journal of Mathematics. 7:751-767.
52. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics 5:113.
53. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.
54. Marston CK, Hoffmaster AR, Wilson KE, Bragg SL, Plikaytis B, Brachman P, Johnson S, Kaufmann AF, Popovic T. 2005. Effects of long-term storage on plasmid stability in Bacillus anthracis. Applied and environmental microbiology, 71:7778-7780.
55. Schmidt HA, Strimme K, Vingron M, Von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics, 18:502-504.
56. Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular biology and evolution, 16:1114-1114.
57. Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. Journal of molecular evolution, 29:170-179.
58. Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proceedings of the Royal Society of London B: Biological Sciences, 269:137-142.
59. Weigand MR, Pena-Gonzalez A, Shirey TB, Broeker RG, Ishaq MK, Konstantinidis KT and Raphael BH. 2015. Implications of genome-based discrimination between Clostridium botulinum group I and Clostridium sporogenes strains for bacterial taxonomy. Applied and environmental microbiology, 81:5420-5429.
60. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one, 5:e9490
61. Kendall M, Colijn C. 2016. Mapping phylogenetic trees to reveal distinct patterns of evolution. Molecular biology and evolution, 33:2735-2743.
62. Jombart T, Kendall M, Almagro-Garcia J, Colijn C. 2017. TREESPACE: Statistical exploration of landscapes of phylogenetic trees. Molecular ecology resources, 17:1385-13-92

CHAPTER 4

NOVEL, PATHOGENIC CLONAL COMPLEXES OF *E. COLI* ALONG A RURAL TO URBAN GRADIENT IN NORTHERN ECUADOR

4.1 Summary

Pathogenic *Escherichia coli* is a leading cause of diarrhea. In South America, particularly in Ecuador, previous epidemiological studies have reported a significant association of *E. coli* strains with diarrhea, yet the population structure and virulence content of *E. coli* populations in this region remain poorly studied. To provide insights into the latter issues, we used molecular pathotyping and whole genome sequencing of 279 *E. coli* strains belonging to six pathotypes (DAEC, EPECa, EPECt, ETEC, EAEC, EIEC). The strains were isolated over 17 months from individuals enrolled in a case/control study of diarrhea in four sites along an urban to rural gradient of Northern Ecuador. Our results revealed a high genomic diversity among the strains, covering almost the complete *E. coli* phylogenetic tree. Despite the high diversity, however, we identified several DAEC, ETEC and EIEC clonal complexes that were significantly ($p<0.05$) or exclusively associated with diarrhea relative to control samples and encoded the hallmark virulence factors (VFs) of the pathotype, suggesting that they might have caused small-scale outbreaks in both rural and urban settings. Notably, several of these

complexes represented DAEC strains, which are generally thought to not be as virulent as strains of other pathotypes. Comparison of VF content revealed that urban strains harbor, in general, more virulence genes, but not antibiotic resistance genes (ARGs), than rural strains and that this observation was mainly driven by two pathotypes, DAEC and EPECa. Collectively, our results revealed novel genomic diversity of *E. coli* related to diarrheal disease in Ecuador, and plausible explanations for the stronger association of certain recovered isolates with disease based on their genomic makeup.

4.2 Introduction

Escherichia coli (*E. coli*) is a gut commensal of vertebrates, including humans (1). This bacterial species plays a critical role in human health, either contributing to metabolism and wellbeing or as causative agent of disease. As a commensal microorganism, *E. coli* is a minor component of the colonic gut microbiome in humans where it typically represents less than 0.1% of the total bacterial cells (~10^8 cells/ml) (2, 3). Nonetheless, this low abundant microbial group contributes to human physiology through the digestion of food, production of vitamin B and K (4), and it has been suggested that it could also have a role protecting against invading enteric pathogens (5). Certain strains (varieties) of *E. coli*, however, can cause a broad range of diseases, including enteric and extra-intestinal infections. Pathogenic *E. coli* strains differ from their commensal counterparts in that they have acquired unique 'fitness' attributes that confer an increased ability to adapt to new niches within the human body and growth, which produces directly or indirectly a broad spectrum of diseases, including diarrhea (6).

Based on the presence of virulence factors, diarrheagenic *E. coli* have been divided into six distinct pathotypes: Enterotoxigenic *E. coli* (ETEC), which produces the heat stable (ST) and/or heat labile (LT) enterotoxins (6,7); Enteroinvasive *E. coli* (EIEC), which harbors plasmid-encoded virulence genes that allow the invasion of intestinal epithelial cells (8), a subset of which form the so called *Shigellae* pathotype (previously thought that it belonged to a different genus *Shigella*); Enteroaggregative *E. coli* (EAEC), identified by its aggregative adherence pattern to cultured HEp-2/HeLa cells and the formation of a biofilm on the gut mucosa (9); Enteropathogenic *E. coli* (EPEC), which induces attaching and effacing lesions on the enterocytes (6, 10); Diffusely

adherent *E. coli* (DAEC), which has been identified by its diffusive pattern of adherence in culture of epithelial HEp-2/HeLa cells and appears to form an heterogeneous group with unknown virulence factors (11); and Shiga-toxin producing *E. coli* (STEC), which produces one or more types of Shiga-toxins (Stx1 and/or Stx2) (12). While a variety of infectious mechanisms can be exploited by these different *E. coli* pathotypes, the development of severe diarrhea in infected individuals is often, but not always, a common symptom. Accordingly, *E. coli* is epidemiologically highly relevant for diarrheal diseases. It has been estimated that, after viral infections, *E. coli* infections are responsible for more than 50% of all diarrheal deaths in children under 5 years old in developing countries (13-15). Of particular concern is the long-term effect of diarrhea in young children with undeveloped gut microbiomes.

In South America, particularly in Ecuador, only a few epidemiological studies have reported association between pathogenic *E. coli* and diarrhea in urban and rural settings (16-18). However, the majority of these studies have been carried out in large cities or rural regions, but no study to date has concurrently evaluated the association of *E. coli* with diarrhea and population structure in urban and rural regions over the same time interval. The increased relocation of people in urban locations (urbanization) have likely increased health risks and challenges associated with the occurrence and prevalence of pathogens causing diarrheal diseases (19). Environmental factors such as poor housing, inadequately treated drinking water supplies, sanitation and waste management in slum areas, coupled with (often uncontrolled) antibiotic usage and spatial proximity in densely populated regions may promote the establishment and rapid spread of highly virulent *E. coli* genotypes in large cities relative to rural areas. How urbanization affects the epidemiology of emerging infectious diseases is unknown, and questions such as whether urban areas promote the establishment of more virulent genotypes than rural regions have been minimally explored (52, 53). Currently, the most rapid growth in urban populations is expected to occur in developing countries (20, 21). Therefore, the characterization of the virulence potential of *E. coli* pathotypes circulating in large cities (≥100,000 inhabitants) and rural areas (≤5,000 inhabitants) is critical in order to evaluate potential risks. A better understanding of the health risk may also lead to improved city planning and surveillance programs that would help decrease the burden of communicable diseases. In Ecuador, no study to date has focused on evaluating whether pathogenic *E. coli* strains circulating in urban areas harbor more, or

perhaps less, virulence determinants than those circulating in rural areas, as well as the occurrence and prevalence of antibiotic resistance genes.

The excretion of enteropathogens by human subjects without diarrhea is a common observation in epidemiological studies (22, 23). Clinical studies have often observed the detection of enteropathogens in healthy individuals not suffering of diarrhea. Several factors, including the genomic makeup of the pathogen, the immune system of the host, and environmental parameters (e.g., the indigenous gut microbiome) might play a role in the outcome. Given that the excretion of enteropathogens in asymptomatic individuals deeply obscures our understanding of the role of those pathogens as etiological agents, it is important to explore whether differences in pathogen genomic content are detectable between strains isolated from individuals whom developed diarrhea versus asymptomatic individuals. In cases of *E. coli* infections, it has not been determined whether or not the virulence content is reduced in asymptomatic individuals excreting *E. coli* pathotypes compared with those individuals who developed diarrhea.

To address these questions, we designed a case/control study of diarrhea (named EcoZUR for *E. coli* en Zonas Urbanas y Rurales) in four sites along an urban to rural gradient of Northern Ecuador in order to evaluate the role that human travel between urban and rural areas played in determining the distribution of pathogenic *E. coli*, and to characterize *E. coli* genome diversity and population structure. Preliminary findings of the EcoZUR study revealed that local travelling to urban areas is a risk factor for diarrheal disease and that the prevalence of different *E. coli* pathotypes can vary across an urban-to-rural gradient (unpublished data). As part of the EcoZUR project, stool specimens collected from individuals enrolled in the study were subjected to PCR screening for the presence of six different *E. coli* pathotypes (DAEC, atypical EPEC, typical EPEC, EIEC, ETEC, and EAEC). Pure *E. coli* cultures from those stool samples that provided positive PCR (+) signal for any of the marker genes tested were subjected to whole genome sequencing (see Materials and Methods). Here, we report on the phylogenomic characterization of those pathogenic *E. coli* strains. Therefore, the aims of this study were to expand the EcoZUR project and use the resulting sequencing data to evaluate the extend of genomic diversity of the selected pathogenic isolates, determine their virulence potential, and characterize their antibiotic resistance profile.

4.3 Materials and Methods

4.3.1 Study design

The genomes analyzed in this study were part of the EcoZUR project. In the EcoZUR project fecal samples from individuals with diarrhea and age-matched controls living in four regions along a gradient of urban to rural areas were taken (Figure 4.1). The four main regions corresponded to 1) Quito (Ecuador's capital) with approximately 1.6 million inhabitants; 2) Esmeraldas, a coastal city in the northwest of Ecuador with 162,000 inhabitants; 3) Borbón, a town in the Esmeraldas Province with ~7,000 inhabitants, and 4) several villages (~150 villages) located along three main rivers: The Cayapas, the Santiago, and the Onzole, with 50 to 500 inhabitants each one. Individuals sampled from Quito and Esmeraldas composed the set of urban specimens while Borbón and the villages composed the set of rural samples. The more urban centers (Quito and Esmeraldas) are densely populated regions with greater access to clean water, sanitation, roads, and medical infrastructure; whereas, the more rural regions (Borbón and rural villages) are less densely populated, with minimal sanitary infrastructure.

Figure 4.1. Geographical map of sampling sites in Ecuador. The map highlights the four territories that comprised the EcoZUR study: Quito, Esmeraldas, Borbón and rural villages. Copyright © Google Maps. 2018. Ecuador, [-2.2995171, -79.7940547,6z].

Individuals of all ages were recruited between April 2014-September 2015 from the Ecuadorian Ministry of Health hospitals and/or clinics of each location (Centro de Salud N°4 Chimbacalle in Quito, Hospital Delfina Torres de Concha in Esmeraldas, and Hospital Básico de Borbón in Borbón). Participants from the rural villages were recruited either through scheduled Ministry of Public Health clinical visits to these rural areas or at Borbón Hospital if they visited the hospital for medical attention. Individuals were recruited for the study if they presented diarrhea (self-reported), defined as three or more loose stools in a 24-hour period and not antibiotic intake in the previous 7 days of the sampling day. Control subjects were recruited if they visited the same medical facility with a non-diarrheal illness, and not vomiting or antibiotic intake in the previous week (7 days). Oral consent for participation was obtained at both the village and household levels. Permission for the study and approval of human subjects was obtained from the Ecuadorian Ministry of Public Health (MSP-DIS-2014-0055-O), Emory Institutional Review Board (IRB) (IRB00065781) and the Universidad San Francisco de Quito (USFQ) Ethical Committee (2013-145M). All participants completed an electronic survey about patient/family demographics, socioeconomic status, medical history, WASH practices, and travel history. Additionally, participants were asked to provide a stool sample for enteric pathogen analysis. Upon receipt of the stool sample, staff members filled two cryoconservation tubes with fecal material and stored them immediately in a liquid Nitrogen dewar for transport to the USFQ laboratory in Quito, where the tubes were transferred to a -80ºC freezer for storage prior to analysis. A total of 907 subjects enrolled successfully completed the survey and 85% of those (771) provided a stool sample. In addition, fresh stool samples were immediately tested for rotavirus antigens using the RIDA Quick Rotavirus test (r-biopharm, Darmstadt, Germany).

4.3.2 Growth conditions and strain pathotyping

For each stool sample, five lactose-positives colonies were isolated on MacConkey's agar media (MKL) and non-lactose fermenting isolates were cultured and tested on Chromocult agar media (Merck, Darmsladt, Germany) (CC) for β-glucoronidase (MUG) activity. Colonies unable to ferment lactose were identified by biochemical tests as *Shigella* or *E. coli* using the API 20E test (BioMérieux, Marcy l'Etoile, France). The five colonies were pooled, re-suspended in 300µl of sterile distilled water, boiled for 10 minutes to release the DNA, and the resulting supernatant was used for PCR testing. Singleplex PCR assays were used on a set of nine different primers to detect the presence of virulence genes associated with each diarrheagenic *E. coli* pathotype (Appendix B.1). Positive pools for *eaeA* were subsequently tested for *stx1* and *stx2* genes for the differentiation of potential EHEC infections. If a pooled sample tested positive for any virulence factor, then each of the five isolates were re-tested individually to identify the specific isolate carrying the virulence gene.

4.3.3 DNA extraction, sequencing and data availability

DNA from *E. coli* colonies was extracted using the Wizard Genomic DNA Purification kit (Promega). The purity and concentration of the DNA was estimated using a NanoDrop spectrophotometer (Thermo Scientific) and the Qubit 2.0 dsDNA high-sensitivity assay (Invitrogen, Carlsbad, CA). DNA sequencing libraries were prepared using the Illumina Nextera XT DNA library preparation kit (Illumina) according to manufacturer's instruction. After this, libraries were run on a High Sensitivity DNA chip using the Bioanalyzer 2100 instrument (Agilent) to determine library insert sizes. An equimolar mixture of the libraries (final loading concentration of 10 pM) was sequenced on an Illumina MiSeq instrument (School of Biological Sciences, Georgia Institute of Technology), using a MiSEQ reagent v2 kit for 500 cycles (2 x 250 bp paired end run). Adapter trimming and demultiplexing of sequenced samples was carried out using the MiSEQ control software v2.4.0.4. In total 316 *E. coli* isolates tested positive for any of the nine PCR assays and from those, 279 were successfully sequenced. The complete set of strains sequenced in this study has been deposited in the NCBI Sequence Read Archive (SRA) under BioProject ID PRJNA486009.

4.3.4 Read quality control, assembly and gene prediction

Raw reads were initially screened for adaptor sequences using Scythe (24) and trimmed at both 5' and 3' ends based on a PHRED score cutoff of 20 using SolexaQA++ (25). Reads ≤ 50 bp after trimming were discarded. Quality-filtered reads were *de novo* assembled using IDBA-UD with pre-corrections (26) and the percent of contamination and genome completeness was assessed based on recovery of lineage-specific marker genes using CheckM (27). Protein-coding sequences were predicted using MetaGeneMark (28), and 16S rRNA gene sequences were identified using barrnap 0.6 (https://github.com/tseemann/barrnap). All predicted genes from the assemblies were taxonomically annotated using MyTaxa (29) and the taxonomic distributions of adjacent genes (in windows of 10 genes) in the concatenated assembly were inspected for possible contamination through barplots named '*MyTaxaPlots'*. The above-described methods and scripts for read quality control, assembly and gene prediction were used as part of MiGA (Microbial Genomes Atlas), a system developed in our laboratory for data management and processing of microbial genomes and metagenomes (http://microbial-genomes.org/) (30).

4.3.5 Bioinformatics inference of Warwick Multilocus sequence type

Sequence type (ST) was bioinformatically determined using the Warwick MLST scheme (31), a genotyping approach relying on the nucleotide diversity of seven housekeeping genes: *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*. The Warwick MLST database was downloaded from the Enterobase website (http://enterobase.warwick.ac.uk/species/ecoli) and the complete set of predicted protein-coding genes were searched against the database using the BLASTn algorithm (32). Alleles were assigned to a ST if the BLASTn hit had ≥ 80% nucleotide identity and ≥ 80% query coverage. In most cases, the sequences matched the alleles in the database with 100% nucleotide identity and 100% coverage. STs were assigned based on the allele combination of the seven genes, as described previously (31). A minimum spanning tree with the sequence types assigned to each bacterial isolate was built using BioNumerics software v7.6.1 (Applied Maths, Sint-Martens-Latem, Belgium). Clermont

phylogroup membership was determined based on the correspondence between Warwick sequence type number and triplex PCR genotype as described in (31).

4. 3.6 Pan-genome and phylogenetic reconstruction

Predicted-coding genes from all bacterial genomes were compared using the scripts *rbm.rb* and *ogs.mcl.rb* from the Enveomics collection (33) to identify shared reciprocal best matches in pairwise genome comparisons and extract the unique core genome (orthologs present in all genomes in one copy). Core orthologous genes were aligned using MUSCLE v3.8.31 (34), with default parameters. The script *Aln.cat.rb* from the Enveomics collection was used to concatenate the multiple alignments into a single file. Phylogenetic tree was constructed using FastTree version 2.1.7 (35) with 1,000 bootstrap replicates and the GTR-GAMMA substitution model for nucleotides. The core-genome-based phylogenetic tree was cross-referenced with metadata and visualized using iTOL (36).

4.3.7 Virulence, antibiotic resistance profile and IncN-type plasmid characterization

The Virulence Factors Database (VFDB) (37) was used to detect virulence genes in our collection of *E. coli* genomes. To avoid limitations of the assembly process such as gaps or truncated genes and miss-assemblies, virulence gene content was assessed using two approaches: first, by BLASTp searches of predicted ORFs against the VFDB using a cutoff ≥ 40% amino acid identity at least 80% coverage of the subject sequence length; and second, by recruiting high-quality reads against the virulence genes on the VFDB and determining gene presence/absence by the number of reads recruited (or not) on the gene. Predicted genes involved in antibiotic resistance mechanisms were identified with BLASTp searches against the antibiotic resistance database (CARD) (38) using a detection threshold of at least 40% amino acid identity at least 80% of the subject sequence covered by the BLASTp alignment. Putative plasmid contigs for each *E. coli* library were identified and assembled using plasmidSPAdes (39). Plasmid annotation based on incompatibility (Inc) groups was performed with PlasmidFinder v1.3

(40) using the *Enterobacteriaceae* database as a reference. Incompatibility groups were assigned based on a cutoff of ≥ 95% nucleotide sequence identity.

4.3.8 Statistical Analysis of Association

All analysis in association between *clinical status*, *lifestyle* or *pathotype* with bioinformatically determined phylogroups was completed using R Studio Statistical Software (http://www.rstudio.org/). Unadjusted odds ratios and corresponding *p*-values were computed using the Pearson's Chi Square Test. If any expected counts were less than five, the Fisher's Exact Test was used instead. Significant associations were determined using an alpha level of 0.05 for all *p*-values in addition to assessing Wald's 95% confidence intervals for the inclusion of the null.

4.4 Results and discussion

4.4.1 Characteristics of individuals enrolled in the EcoZUR project

In this study, a total of 279 *E. coli* strains that were PCR positive for at least one of the nine pathogenicity marker genes evaluated (see methods and materials) were isolated from stool specimens from individuals with diarrhea and controls living in urban and rural areas in Northern Ecuador and sequenced. We comparatively studied the population structure of these strains in three main categories: 1) *clinical status*, which denoted whether the isolate was obtained from an individual with diarrhea or an individual excreting pathogenic *E. coli* but who had not developed diarrhea (we refer to the latter as asymptomatic samples or infections), 2) *lifestyle*, which corresponded to whether the isolate was obtained from an individual living in a urban or rural region (see Materials and Methods for details), and 3) *pathotype*, which corresponded to the PCR detection of any of the markers genes that were used to discriminate among different *E. coli* pathotypes.

To determine if any host-associated biases existed in the collection of isolates that were included in this study, we first evaluated the demographic information of the sampled individuals. We specifically evaluated if significant differences related to age and gender existed between urban and rural sites. Our cohort of individuals was mostly

composed by young children between 0-24 months of age at both sites. No statistical differences in age category or gender were observed between individuals living at urban or rural sites (Table 4.1). Therefore, we ruled out any potential bias in population structure of *E. coli* strains that were introduced by demographic structuring of the hosts.

Table 4.1. **Demographic information of individuals enrolled in the EcoZUR project whose stool specimens where PCR-positive for pathogenic *E. coli* gene markers.**

	Urban (n = 131), n (%)		Rural (n = 128), n (%)		Total (n=259), n(%)	*pval*
Gender						
Male	74 (56.4)		70 (54.6)		144 (55.6)	
Female	57 (43.5)		58 (45.3)		115(44.4)	
Age (months)	**# counts**	**mean (SD)**	**# counts**	**mean (SD)**		
0-24	49	14.2 (6.3)	49	14.0 (4.9)	98 (37.8)	*0.54*
25-60	18	41.2 (11.3)	30	38.5 (10.9)	48 (18.5)	*0.42*
61-180	27	115 (31.4)	21	119 (35.6)	48 (18.5)	*0.85*
>181	37	455 (195.6)	28	491 (213)	65 (25.1)	*0.42*

From the set of 279 isolates that were successfully sequenced in total, 168 were obtained from cases of diarrhea and 110 were detected from control individuals, which indicated a state of asymptomatic infection. 139 strains were isolated from individuals living in urban regions (Quito or Esmeraldas) while 140 were cultured from people in rural settings (Borbón or rural villages). DAEC was the most abundant pathotype found in the dataset (n=99), followed by atypical EPEC (n=66), ETEC (n=48), EAEC (n=45), EIEC (n=15), and typical EPEC (n=4) (Appendix B.2).

4.4.2 Genome assembly and quality checking

The assembled genomes consisted of 330 contigs on average, with a G+C% content of 50.7% and an estimated genome size of 4.9Mb. They were predicted to contain, on average, a total of 4,894 putative protein-coding genes sequences, 7 rRNA operons and 86 tRNA genes. In addition, the average coverage of the genomes sequenced in this study was 124X (min=7X, max=1,731X). To determine the quality of our assembled genomes, we estimated the completeness and contamination percent of

the assembled genome and manually inspected the taxonomic affiliation of adjacent genes, in windows of 10 genes, using *MyTaxaScan* plots (30) (see Materials and Methods). The estimated average percent of completeness and contamination for our dataset were 98.5% (SD±1.9) and 0.97% (SD±0.9), respectively. However, for 16 isolates a substantial signal of genomic contamination with *Klebsiella pneumoniae* (10 to 50% of the total genome size, depending on the genome considered) was detected including universal genes that are usually found in single copy and are not often subjected to lateral transfer. Therefore, we concluded that these isolates were most likely contaminated by *K. pneumoniae* DNA instead of displaying a signal of horizontal gene transfer (HGT) and therefore, were discarded from further analysis.

4.4.3 Bioinformatic detection of *E. coli* marker genes used for PCR-based pathotyping

To evaluate if the PCR-based pathotyping was consistent with genome content recovered from the assembly of each isolate, we performed a bioinformatic detection of the panel of virulence genes screened by PCR. We assessed the level of agreement by using a read-based approach in which high-quality reads were recruited against a *E. coli* reference marker genes and the presence/absence of each marker gene was determined by the number of reads recruited (or not) on the gene and the percent of the gene length that was covered (see Materials and Methods). Overall, our results showed a high level of agreement where 81.3% (214 isolates out 263 total) of the genomes showed consistent results between PCR and bioinformatic pathotyping. Only 7 isolates showed disagreement (2.6%) where bioinformatic analysis detected a different gene than the one detected with PCR. In 42 isolates (16%) no marker gene was detected with the bioinformatic approach, indicating a false-positive PCR amplification. However, in 7 isolates, we detected the presence of an additional marker gene that was not identified through PCR (Appendix B.2).

4.4.4 Multilocus sequence analysis (MLSA)

To perform an initial screening of the genetic diversity in our *E. coli* dataset, we carried out an *in silico* Multilocus Sequence Analysis (MLSA) characterization based on the allelic diversity of seven housekeeping genes included in the Warwick *E. coli* MLST

scheme (*adk, fumC*, *gyrB, icd, mdh, purA* and *recA*) (31). We used a minimal spanning tree (Figure 4.2a) to visualize the diversity and clonality of the assigned sequence types (ST) to our genomes. Overall, our results revealed a high genetic diversity in the strains circulating in urban and rural areas in Northern Ecuador (Appendix B.3). We found a total of 84 unique STs, from which 26 STs (31%) were equally distributed between urban and rural areas, 22 STs (26%) were specific for urban areas, and 36 STs (43%) were found only in rural regions. The top five most prevalent STs found in our dataset were ST10 (n=34), ST38 (n=15), ST131 (n=14), ST4 (n=11), and ST270 (n=8). This distribution was consistent with previous MLSA studies showing that ST10 is one of the most frequently found STs in human fecal samples and in food samples (43).

In general, ST270, ST4, ST131, and ST38 were more prevalent in cases of diarrhea while ST10 was more prevalent in asymptomatic individuals (Figure 4.2b). When sequence type assignations were evaluated for *lifestyle*, we observed that ST10 was equally distributed between urban and rural areas (17 rural and 17 urban), while ST38 and ST131 were more abundant in urban areas and ST4 and ST270 were more prevalent in rural regions (Figure 4.2c). When evaluated by *pathotype*, we observed that DAEC strains were assigned to ST131 and ST38 mainly, ETEC was restricted to one ST (ST4) and EIEC was mostly assigned to ST270 (Figure 4.2d). We also found 25 isolates presenting 22 new alleles combinations that have not been described previously in the reference database as August 2018 (Appendix B.3). Of particular interest was the observation that ST131, a highly dispersed and potentially virulent group of *E. coli* strains that has been linked to a rapid global increase and prevalence of antimicrobial resistance, was found to be among the most prevalent STs in our dataset and present mostly in individuals living in urban regions. A recent study by Chiluisa-Guacho 2018 (44, 45) also reported the detection of several B2-ST131 clones isolated from patients diagnosed with urinary tract infections (UTI) in Quito-Ecuador, which appear to have been disseminated in hospitals and community settings. This finding highlights the need to maintain a surveillance system for the detection of *E. coli* clones that might represent a public health treat, particularly those that are globally disseminated.

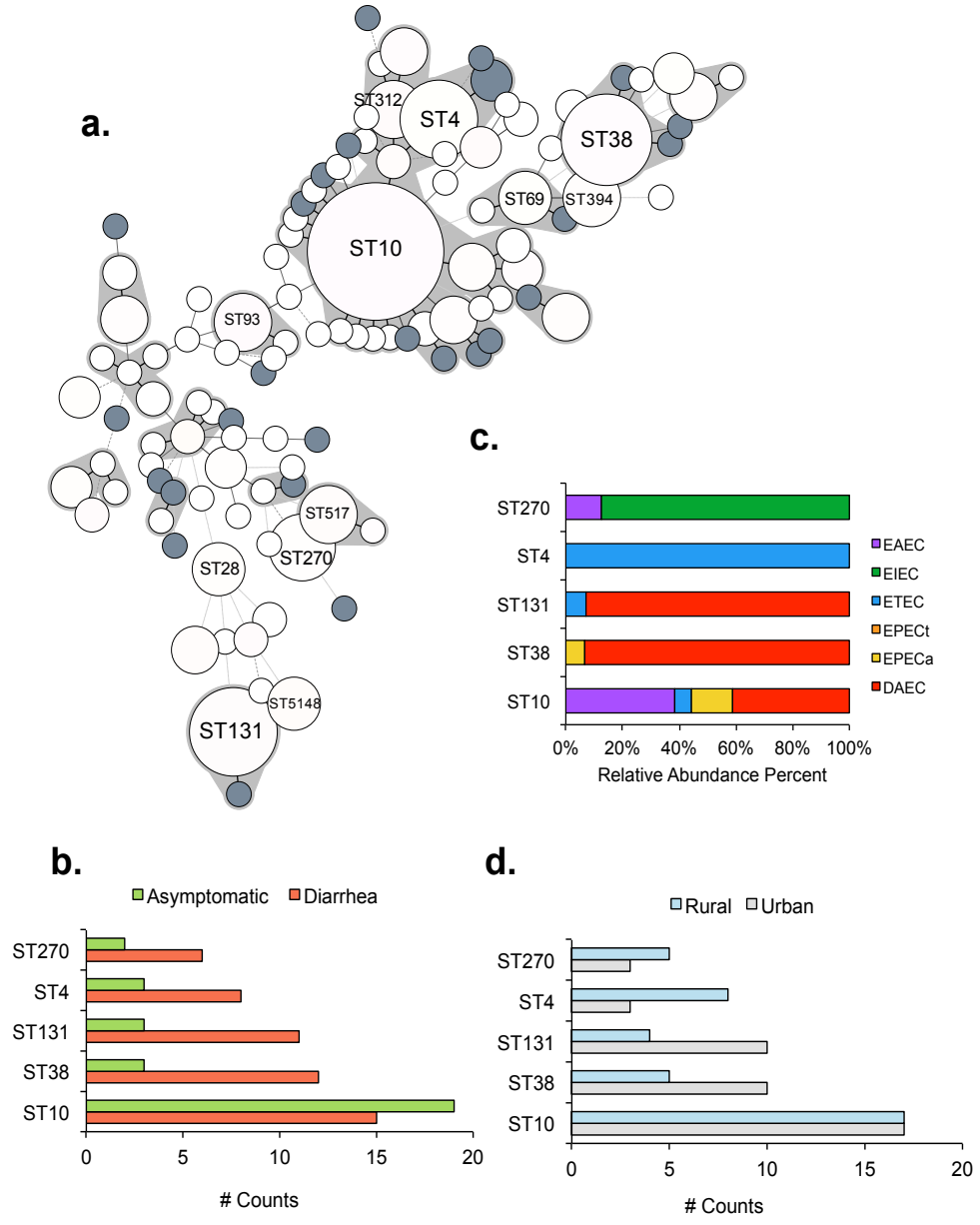


Figure 4.2. MLSA profile of pathogenic *E. coli* isolates circulating in Northern Ecuador. Panel a shows a minimum spanning tree based on the combination of allelic variants of the Warwick MLST profile. Each node represents a single sequence type (ST) and the size of the node is proportional to the number of isolates that belong to a given ST. The length of the branch between each node is proportional to the number of distinct alleles that differ between the two linked nodes. Nodes in dark grey represent new alleles combinations not observed in the Warwick MLST database. Only nodes with ≥ 5 strains were labeled with the corresponding sequence type number. Panels b and d

show the prevalence of the top five most prevalent STs detected in our *E. coli* dataset per *clinical status* and *lifestyle*, respectively. Panel c shows the relative abundance of each *pathotype* in the top 5 most abundant STs.

4.4.5 Population structure of the *E. coli* isolates

To gain insights into the evolutionary relationships of *E. coli* pathotypes circulating in Northern Ecuador, we characterized the pangenome of the isolates in our dataset and estimated the phylogenetic relationships based on the core genome. The pangenome was composed by 17,596 orthologous genes (OGs) in total, from which 1,875 were part of the strict core genome (present in all strains) and 3,661 were observed in at least 90% of the strains. 13,935 OGs were part of the accessory genes. We used core OGs present in all strains with exactly one gene per genome to build a maximum likelihood (ML) phylogeny and determine whether evidence of population structure by *clinical status* (case/asymptomatic), *lifestyle* (urban/rural) or *pathotype* (EAEC, EIEC, ETEC, EPECt, EPECa, DAEC) could be detected. Phylogroup membership was assigned based on the corresponding ST as described in (31) and confirmed by visualizing branching patterns in the ML tree. To evaluate the *pathotype* category, we only screened those strains for which the pathotype diagnostic gene was detected both experimentally and bioinformatically (n=224).

Initial phylogenomic characterization showed that the total set of strains was distributed in five main clades: A (115 strains), B1 (68 strains), D (49 strains), B2 (30 strains) and E (one strain) (Figure 4.3a). Only one strain (E5.5) belonging to clade E was observed in our dataset and a single isolate (B37.6) fell into the cryptic clade that branches outside of the established *E. coli sensu stricto*. These observations were consistent with previous studies in South America (French Guiana, Colombia and Bolivia) reporting that strains isolated from this tropical region belonged mainly to the phylogenetic groups A (55%) and B1 (21%), whereas strains from the phylogroups D (14%) and B2 (10%) were less common (46, 47).

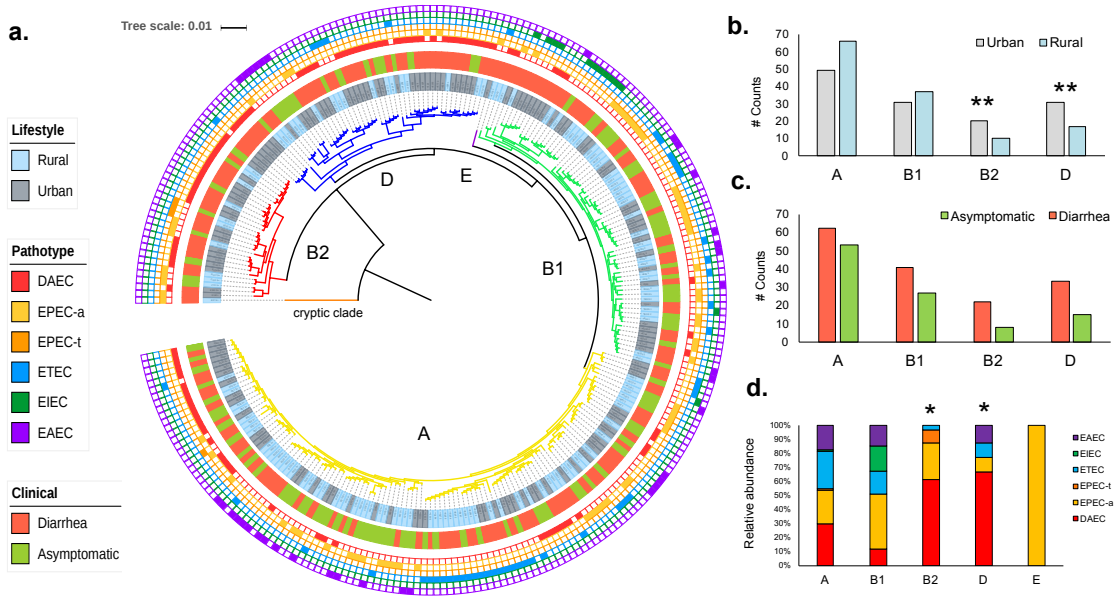


Figure 4.3. Phylogenetic relationships and population structure of the 263 *E. coli* isolates. Panel a shows a phylogenetic reconstruction calculated from the concatenated alignment of 1,200 core orthologous genes using FastTree 2.1.7 with the GTR model for nucleotide evolution and 1,000 SH-like local support replicates. The tree is also cross-referenced with metadata in three categories: 1) *lifestyle* (urban, rural), 2) *clinical status* (diarrhea, asymptomatic) and 3) *pathotype* (DAEC, EPECa, EPECt, ETEC, EIEC and EAEC). Panels b, c and d show the relative abundance of each one of the above-mentioned categories by phylogroup. Asterisks denotes those phylogroups for which significant associations with each tested category were found.

Initial characterization by *lifestyle* showed that clades A and B1 had higher prevalence of isolates from individuals living in rural areas while clades B2 and D contained more isolates circulating in urban areas (Figure 4.3d). Accordingly, significant associations were detected between clade D (OR = 2.05, 95% CI = 1.02-4.12, *p=0.04*) and B2 (OR = 2.37, 95% CI = 1.08-5.21, *p=0.03*) with urban isolates (Figure 4.3a, Table 4.2). In other words, we found that *E. coli* pathotypes present in phylogroups B2 and D were twice as likely to be associated with urban than rural subjects. The genetic

structure of both commensal and pathogenic *E. coli* strains is thought to be shaped by multiple host and environmental factors. As it has been previously illustrated, host characteristics such as diet, gut morphology, microbiome structure and body mass seem to be important predictors of the distribution of *E. coli* phylogenetic groups (48-49). It is possible that socio-demographic factors such as dietary habits and sanitation practices might have contributed to the population structure observed in our collection of genomes although a sampling artefact should also be considered. However, our initial screening of the demographic composition of the host (individuals) showed no significant differences by gender or age group, in both rural and urban groups. Roughly, an equal representation of rural and urban genomes was included in this study (Table 4.1), which support the observed differences in genetic structure by host lifestyle.

Table 4.2. Association of the Clermont phylotypes with urban vs. rural categories. Note that *E. coli* isolates affiliated with phylogroup B2 and D were twice as likely to be from urban samples.

Clermont	**Overall**	**Urban**	**Rural**	**OR (95% CI)**	***p-value***
A	94 (42.0)	44 (36.1)	50 (49.0)	0.59 (0.34, 1.00)	0.05
B1	47 (21.0)	20 (16.4)	27 (26.5)	0.54 (0.28, 1.04)	0.07
B2	35 (15.6)	25 (20.5)	10 (9.8)	2.37 (1.08, 5.21)	***0.03****
D	44 (19.6)	30 (24.6)	14 (13.7)	2.05 (1.02, 4.12)	**0.04***
Shigella	4 (1.8)	3 (2.25	1 (1.0)	2.55 (0.26, 24.9)	0.63

Statistical analysis by *pathotype* also showed a non-random distribution of pathogroups in the phylogenetic tree (Table 4.3). DAEC was most likely to be found in clade A, B2 and D ($p<0.001$) while EAEC and ETEC were both most likely found on phylogroup A ($p=0.008$ and $p<0.001$, respectively). EPECa and EIEC were predominately found on clade B1 ($p<0.001$ in both groups) while EPECt was only found in clades A and B2 but not significant association was detected. Analysis by *clinical status* however, showed that all strains obtained from either cases of diarrhea or asymptomatic samples were

distributed along all five phylogroups and not significant association was detected in any phylogenetic group (Figure 4.3b, Table 4.4). Overall, our results suggested that a degree of population structuring exits in our collection of *E. coli* isolates, which is mainly driven by *pathotype* and host *lifestyle* but not by diarrhea disease.

Table 4.3. Distribution of *E. coli* pathotypes in the Clermont phylogroups.

E. coli Pathotypes	Overall Prevalence n (%)	Clermont Phylogroups A n (%)	B1 n (%)	B2 n (%)	D n (%)	Shigella n (%)	*p-value*
DAEC	**88 (39.3)**	**31 (35.2)**	**3 (3.4)**	**20 (22.7)**	**34 (38.6)**	**0 (0.0)**	**<0.001**
EAEC	46 (20.5)	26 (56.5)	12 (26.1)	0 (0.0)	8 (17.4)	0 (0.0)	0.008
EPECa	**38 (17.0)**	**9 (23.7)**	**17 (44.7)**	**10 (26.3)**	**1 (2.6)**	**1 (2.6)**	**<0.001**
ETEC	**34 (15.2)**	**25 (73.5)**	**7 (20.6)**	**1 (2.9)**	**1 (2.9)**	**0 (0.0)**	**<0.001**
EIEC	**13 (5.8)**	**1 (7.7)**	**8 (61.5)**	**1 (7.7)**	**0 (0.0)**	**3 (23.1)**	**<0.001**
EPECt	5 (2.2)	2 (40.0)	0 (0.0)	3 (60.0)	0 (0.0)	0 (0.0)	0.07
Total	224	94	47	35	44	4	

Table 4.4. Association of the Clermont phylogroups with diarrheal disease status. No significant differences were observed for isolates of different phylogroups with diarrhea vs asymptomatic samples.

Clermont Phylogroup	**Overall** ***n (%)***	**Case** ***n (%)***	**Control** ***n (%)***	**OR (95% CI)**	***p-value***
A	94 (42.0)	52 (38.5)	42 (47.2)	0.70 (0.41, 1.21)	0.20
B1	47 (21.0)	26 (19.3)	21 (23.6)	0.77 (0.40, 1.48)	0.44
B2	35 (15.6)	24 (17.8)	11 (12.4)	1.53 (0.71, 3.31)	0.27
D	44 (19.6)	29 (21.5)	15 (16.9)	1.35 (0.68, 2.69)	0.39
Shigella	4 (1.8)	4 (3.0)	0 (0.0)	--	0.15

4.4.6 Resistome of *E. coli* isolates

To evaluate the antimicrobial resistance profile of our set of *E. coli* isolates and determine which are the most abundant antibiotic resistance genes (ARGs) encoded in their genomes, a gene content analysis was performed using as reference sequences 1,371 genes included in the Comprehensive Antibiotic Resistance Database (CARD) (38). Presence of a reference genes was defined as a BLASTp match of ≥ 40% percent identity and ≥ 80% of the query length being covered (aligned). Detected ARGs were grouped in nine classes according to the antibiotic they confer resistance to, and their relative frequency of the classes was examined based on three categories as above: *clinical status*, *lifestyle*, and *pathotype*.

Our results showed that fluoroquinolones were the most abundant group of ARGs found in the collection of genomes, followed by animocoumarins, and polymyxins. Fluoroquinolone class, which include antibiotics compounds used to treat a variety of common illnesses such as respiratory and urinary tract infections (ciprofloxacin, gemifloxacin, levofloxacin, among others) included a total of nine genes, while eleven genes composed the animocoumarin class and four genes composed the polymyxin class. No significant differences were observed when the *E. coli* resistome was compared among isolates by *clinical status*, *lifestyle* or *pathotype*. However, we observed a trend of increased abundance of beta-lactamases and macrolides in urban areas compared to rural areas, which did not reach statistical significance (Figure 4.4).

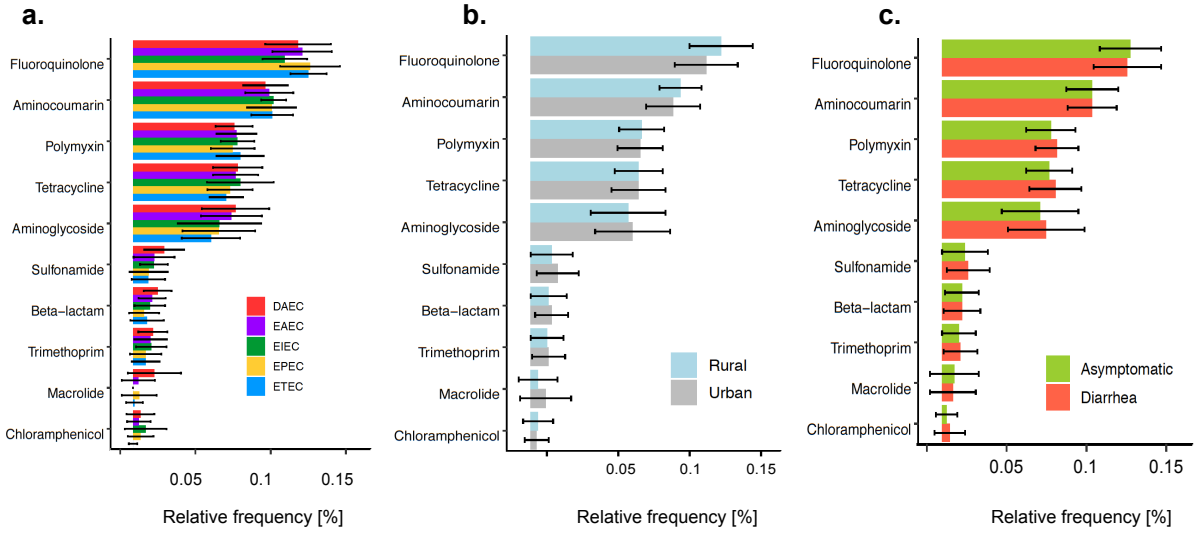


Figure 4.4. Antibiotic resistance gene profile of *E. coli* isolates. Barplots show the average frequency of ARGs in *E. coli* isolates categorized by antibiotic class to which they confer resistance. Panels a, b, c show relative frequency of ARG classes discriminated by *pathotype*, *lifestyle* and *clinical status* categories respectively.

4.4.7 Virulence content of pathogenic *E. coli* strains circulating in Northern Ecuador

To assess the virulence profile in our *E. coli* isolates and determine if isolates from urban subjects' harbor, in general, more virulence genes than their rural counterparts, we performed a gene content analysis evaluating the presence and distribution of 293 experimentally validated virulence factors (VF) specific for *E. coli* that were retrieved from the Virulence Factors Database (VFDB). In total, we found 245 out of 293 virulence genes in our collection of genomes, which represented ~86% of the total database. The analysis of the distribution of the virulence genes showed that more than 90% of the isolates harbored a core set of 32 virulence factors. These genes were functionally annotated in four categories: *iron acquisition*, which included gene clusters associated with the biosynthesis of siderophore systems comprising ferrienterobactins and aerobactins; *motility*, which included genes encoding for several fimbrillin subunits of the common pilus; *stress survival,* which was represented by the *ompA* gene and *type III secretion system effectors* including *espL1*, *espL4* and *espX5* (Figure 4.5).

We also found that strains circulating in urban areas encoded, in general, more VF than strains circulating in rural areas (*p=0.001*) (Figure 4.6a). This overall pattern was driven by two pathotypes mainly: DAEC (p=0.04) and EPECa (p=0.009) (Figure 4.6b). Analysis for the remaining pathotypes (EAEC, EIEC, EPECt or ETEC) showed no significant differences between rural and urban strains. However, of particular interest was the observation that EPECt strains had consistently higher virulence content than any other pathotype. Analysis of virulence potential evaluated by *clinical status*, showed not significant differences between strains isolated from cases of diarrhea and asymptomatic (Figure 4.6c and 4.6d). However, for a few clonal complexes associated mostly or exclusively with diarrheal samples, high VF content was noted compared to their closest related clonal complexes that showed no strong association with disease.

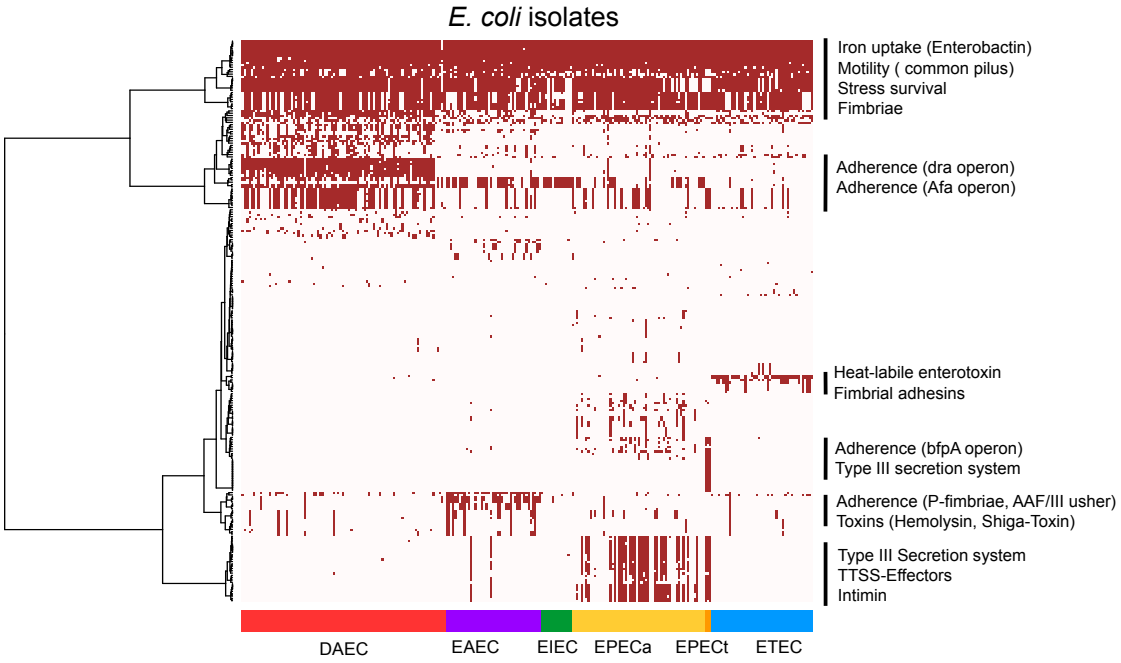


Figure 4.5. Virulence gene content of *E. coli* isolates. Heatmap shows a hierarchical clustering based on the presence (dark brown) or absence (light pink) of known *E. coli* virulence factors (n=293). In total, 245 out of 293 virulence genes were detected in at least one genome, which represented ~86% of the total database. Note that a core set of virulence genes were detected in >90% of the collection of strains. These genes were functionally annotated in four categories: *iron acquisition*, *motility*, *stress survival* and *P-fimbriae.*

The distribution of virulence genes along the core genome phylogeny and specifically per pathotype were also evaluated. For this, all virulence genes detected in our collection of isolates were categorized in six main categories according to the classification found in the VFDB: iron uptake, adherence, type III secretion system (TTSS), toxins, effectors and others. Our results revealed that genes encoding for systems related to acquisition of environmental iron were the most frequently found in all pathotypes. In DAEC the second most frequently found category was adherence whereas in EPECa it was type III secretion system (TTSS). The latter category was significantly more abundant in isolates from urban areas than isolates from rural areas. In ETEC, the second most abundant category was effectors; however, significant

differences were observed for TTSS, with isolates from rural areas carrying, in general, more TTSS genes than isolates from urban settings. Finally, in EAEC group, the second most abundant category was adherence. Diarrhea-associated ETEC isolates had significantly more toxin genes than isolates from asymptomatic individuals.

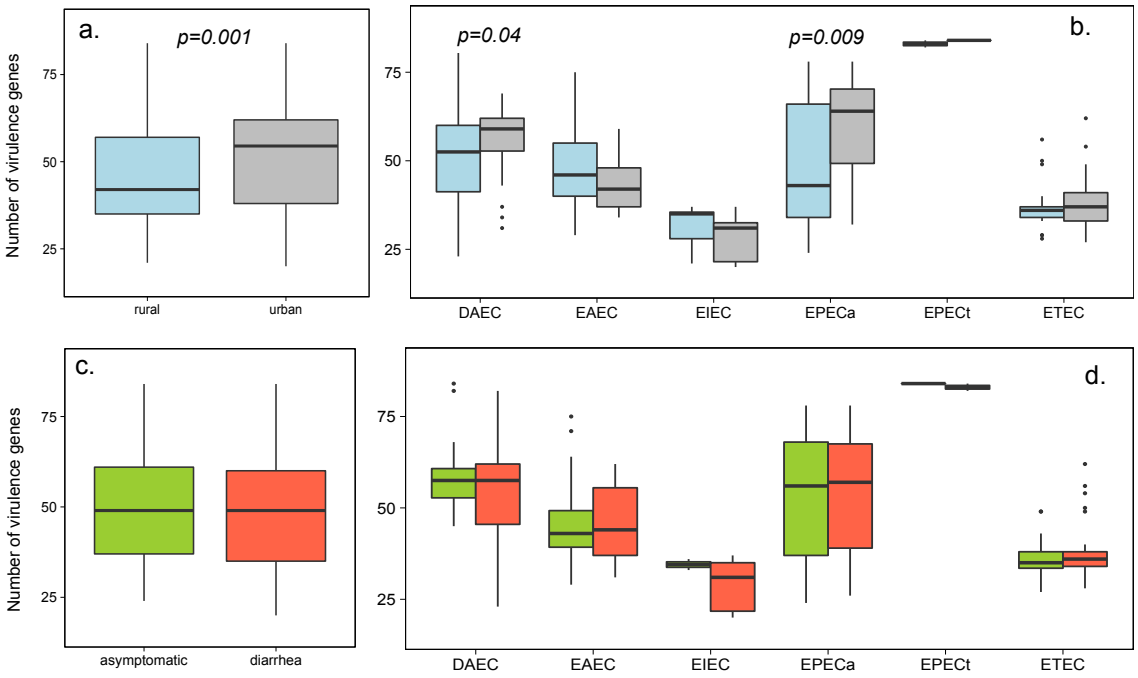


Figure 4.6. Distribution of virulence genes in *E. coli* isolates by lifestyle and clinical status. Boxplots shows the distribution of the number of detected virulence genes found in the genome of *E. coli* isolates categorized by lifestyle (urban vs. rural, top left) and clinical status (diarrhea vs. asymptomatic, bottom left). The quantification is further separated by pathotype (right panels). Isolates circulating in urban areas (gray) carry, in general, more virulence genes than isolates circulating in rural areas (light blue). However, note that this overall pattern is driven by two pathotypes mainly: DAEC and EPECa.

4.4.8 Detection of *Afa/Dr* adhesion operons in DAEC strains recovered from diarrhea and asymptomatic individuals

Next, we aimed to determine if the virulence potential of our DAEC strains, the most prevalent pathotype in our collection of strains, isolated from asymptomatic individuals was reduced compared with those isolates obtained from individuals who developed diarrhea. For this purpose, we characterized three adhesin operons that have been previously described as key virulence determinants mostly contributing to the pathogenicity of the DAEC pathotype: the *Afa*, *Dra* and *Daa* operons (regarded together as Afa/Dr adhesins) (11). Specifically, we evaluated the frequency of detection of the genes contained in these three operons in diarrhea and asymptomatic individuals. The *afa* operon is composed of seven genes (*afaFABCDGE*). The *afaC* complex was composed by two subunits (I and II) and four types of AfaE have been described in the VFDB database so far. Thus, in total we evaluated the presence of 12 genes composing the Afa operon. Our results showed that the genes *afaF*, *afaA*, *afaB* , *afaC-I*, *afaC-II* and *afaD* were frequently detected in our set of DAEC strains, with at least 80% of the genomes in both diarrhea and asymptomatic groups contained these set of genes (Figure 4.7a).

Conversely, the gene *afaD* was totally absent in our DAEC dataset. Several types of the *afaE* gene were detected in our dataset, which together composed ~80% of the frequency of detection. Although the genes *afaE–I* and *afaE–III* were more frequently observed in DAEC isolates coming from asymptomatic than diarrhea cases, this difference was not statistically significant (Figure 4.7a). Similar results were observed for the Dra (*draABCDPE*) and Daa (*daaAFCD*) operons (Figure 4.7b and 4.7c). Overall, no significant differences in the frequency of detection of these adhesin operons were observed between diarrhea and asymptomatic isolates, which suggests that other factors such as host immune system and genotype or gut microbiota resilience to perturbation rather than pathogen's virulence potential might be contributing to the excretion of enteric pathogens in person without diarrhea.

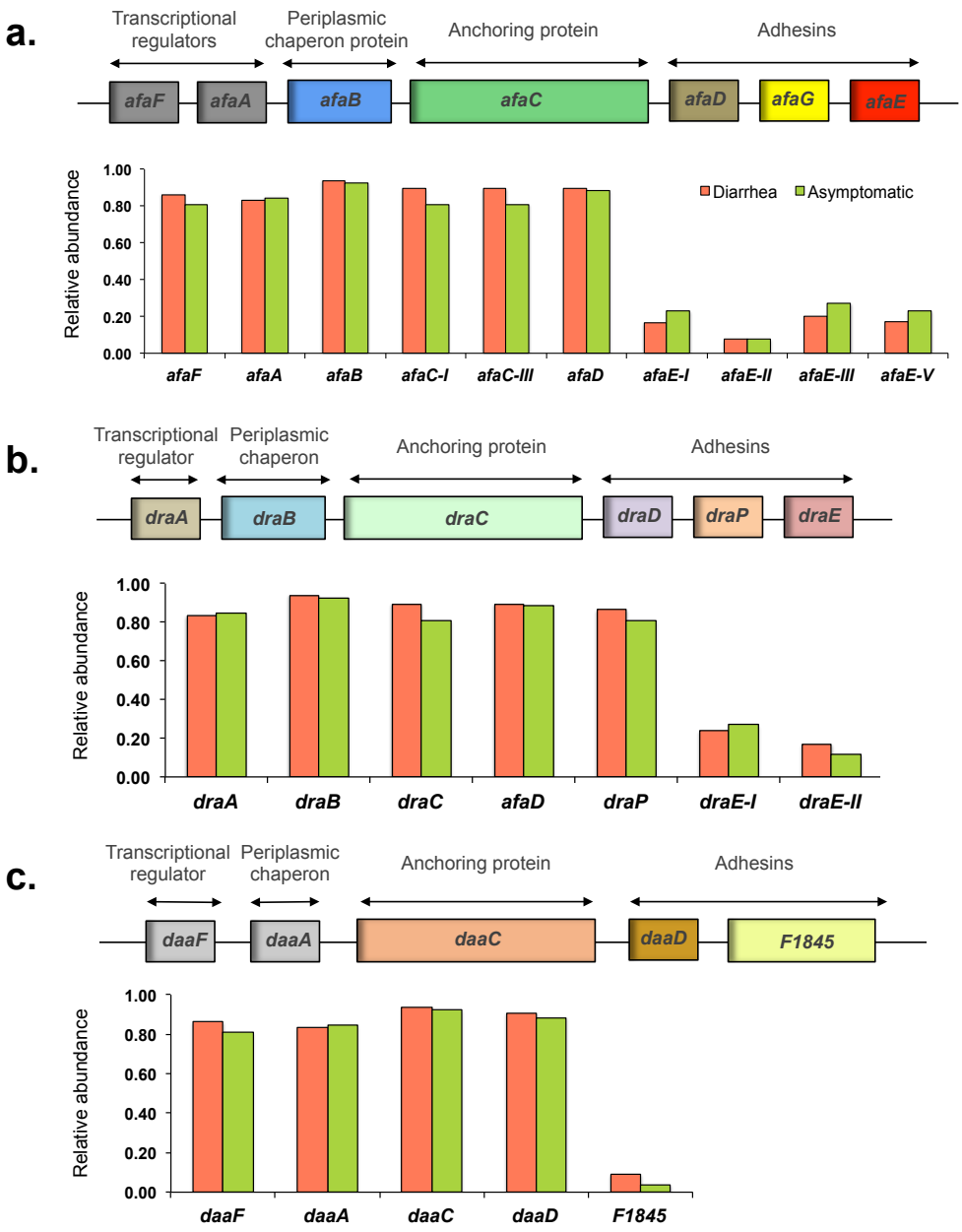


Figure 4.7. Detection of *Afa/Dr* adhesion operons in DAEC isolates recovered from diarrhea and asymptomatic individuals. Panel a shows the relative abundance of the *afaFABCDGE* operon between isolates recovered from diarrheal and asymptomatic samples. Panel b and c correspond to *draABCDPE* and *daaFACD* operons, respectively.

4.4.9 Prevalence of Inc type plasmids and associated resistance genes in *E. coli* isolates

To advance our understanding of the dynamics behind the spread of ARGs that are disseminated through plasmids, we performed a screening of the plasmids that are being carried by the *E. coli* isolates in our dataset. In total, 204 plasmids were retrieved and annotated based on incompatibility type (Inc type) using the plasmidFinder database (40). Plasmid incompatibility is defined as the inability of two related plasmids (same origin of replication) to be stably maintained and propagated in the same cell line due to competition for cellular machinery (41). Currently, 27 Inc groups are recognized in *Enterobacteriaceae* by the Plasmid Section of the National Collection of Type Cultures (London, United Kingdom), including six IncF (FII to VII) and three IncI (I1, Iγ, I2) variants. The analysis showed that a large heterogeneity of plasmid Inc types have been circulating among *E. coli* isolates in Ecuador (Figure 4.8). From these, 61 replicons belonged to incompatibility group *IncFIA* (30%) and 63 to incompatibility group *IncFII* (31%). In addition, 18 plasmids (9%) belonged to *Shigella boydii* group. Within *IncFIA* and *IncFII* groups, the *plasmidF*, *pRSB107*, and *pBS512_2* were the three most abundant plasmids found in our collection and all three were also more prevalent among rural isolates (Figure 4.8a). When classified by clinical status, we found that *plasmidF*, *pHN7A8*, *pRSB107, pMG828-1*, and *pCoo* were more abundant in cases of diarrhea than asymptomatic individuals (Figure 4.8b). When evaluated by pathotype, we observed that plasmids *pMG828* and *pC15-1a* were more prevalent in DAEC strains whereas plasmids *pBS512_2* and *pRSB107* were more abundant in EPECa strains. Plasmid *pCoo* was observed more frequently in ETEC strains while plasmid *pHN7A8* was observed frequently in EIEC strains (Figure 4.8c). However, the distribution of plasmids by pathotype showed that plasmids were not exclusively carried by a single *E. coli* pathotype.

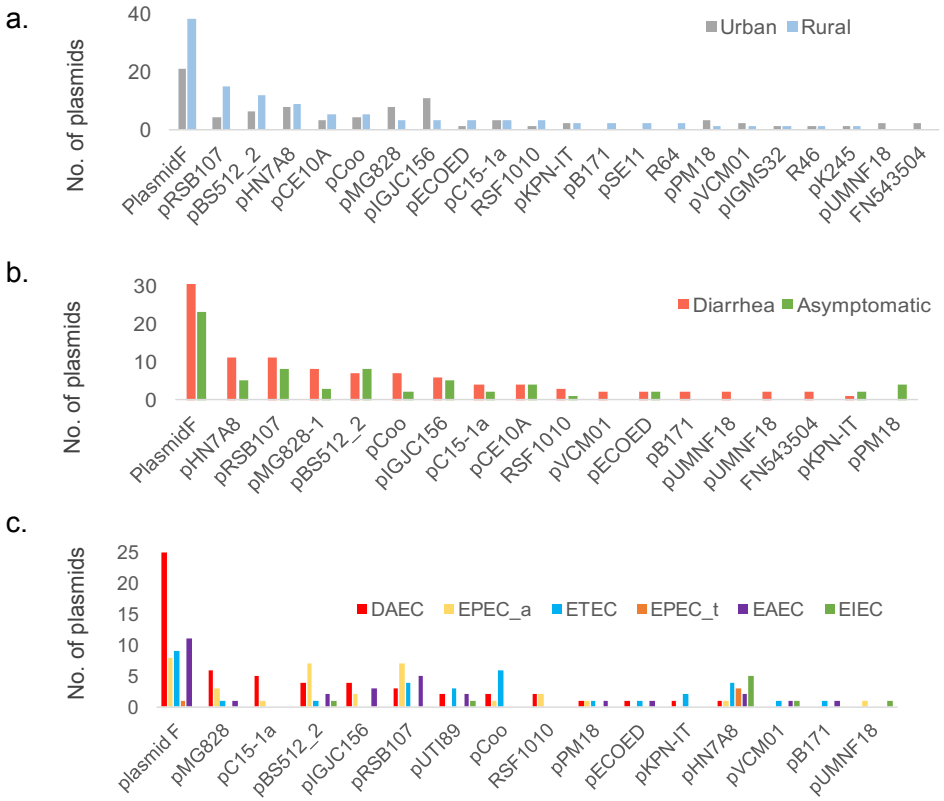


Figure 4.8. Prevalence of Inc type plasmids and associated resistance genes among our *E. coli* isolates. Bar plots show the distribution of the different plasmids retrieved from our collection of isolate genomes. Overall, our results revealed that there was a large heterogeneity of plasmid Inc types circulating in Ecuador.

The genomic content of plasmid *pRSB107*, the most abundant multi-resistant replicon detected in our isolates, was examined more closely. Plasmid pRSB107 encoded nine different antibiotic-resistant determinants, two iron acquisition systems and other putative virulence-associated functions (42). In total, 19 recovered plasmids were annotated as *pRSB107* from our collection and most of them were found in rural communities (15 from rural and 4 from urban areas). These plasmids showed partial recovery of the entire sequence when compared to the reference plasmid *pRSB107* (accession number AJ851089) (Figure 4.9). BLASTn searchers of the predicted genes

from the recovered sequences against the genes from the reference plasmid *pRSB107* showed the presence of genes involved in virulence, iron sequestration and plasmid maintenance (*ugpB*, *ugpA, ugpC, vagD*, *vagC, iucA, iucB*) in eight plasmids. In addition, the plasmid recovered from one strain isolated from an urban area in Quito (Q53) carried genes involved in the resistant to Streptomycin (*strA*, *strB*) and Sulfonamide (*sulII*).

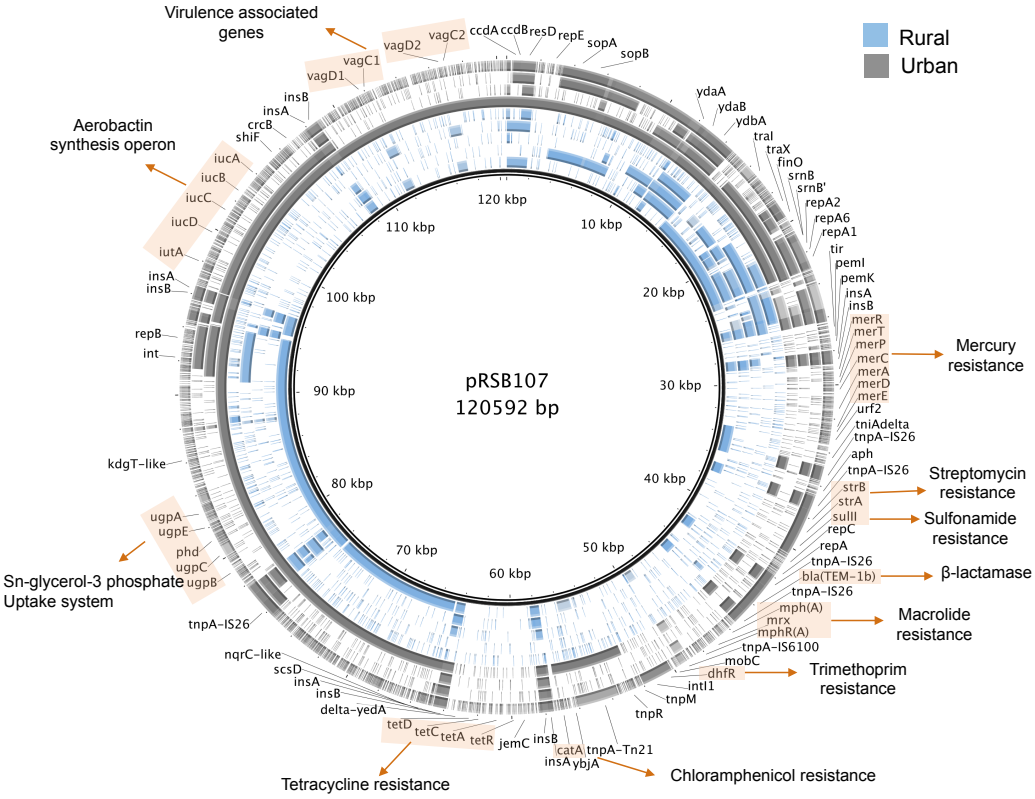


Figure 4.9. Circular plots of the pRSB107 plasmid. Black inner ring: pRSB107 sequence used as reference for the alignment (AJ851089). Blue inner circles represent recovered contigs from *E. coli* isolates from rural areas annotated as pRSB107; grey outer circles represent contigs from urban areas. Genes involved in virulence and antibiotic resistance are highlighted in orange. The plot was constructed using BRIG plots v0.95

Finally, we attempted to evaluate if the phylogeny of *E. coli* isolates based on conserved plasmid genes recapitulates that of the main chromosome. For this purpose, we selected a subset of *E. coli* isolates in which the plasmid F was recovered and compared the tree topology of the plasmid based on the *repA* gene to the core genome tree (Figure 4.10). Comparison of the tree topologies showed inconsistencies between the two phylogenies, revealing that these plasmids are moving among *E. coli* isolates isolated from rural and urban areas and they don't seem to be restricted to a specific host genotype.

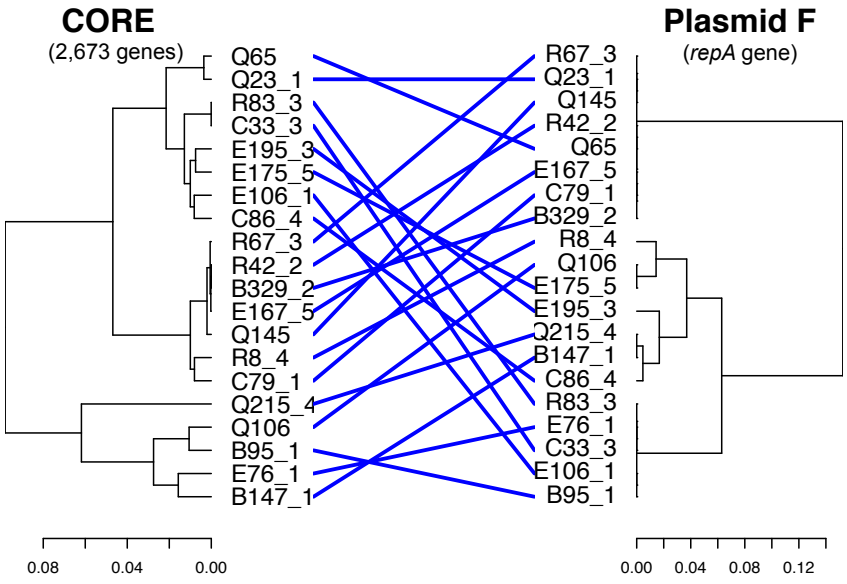


Figure 4.10. Tanglegram comparing tree topologies of maximum likelihood phylogenies of the core genome (left) and the *repA* gene from the plasmid F (right). Crossing blue lines indicate inconsistencies between the two topologies, caused most likely by the horizontal gene transfer of the corresponding plasmids. Phylogenetic trees were constructed using RAxML v8.0.19 with the GTRGAMMA modeling and 1,000 bootstraps.

4.5 Conclusions and recommendations

In this study, we used next generation sequencing data to characterize the extend of genomic diversity and population structure of 279 *E. coli* isolates from individuals diagnosed with diarrhea and age-matched controls living in large cities (urban) and small villages (rural) in Northern Ecuador. We comparatively studied the population structure and virulence potential of these strains in three main categories: *clinical status*, *lifestyle* and *pathotype*. Our major findings revealed a significant, but rather moderate, degree of genetic cohesion (population structure) in some isolates that was driven by pathotype and host lifestyle mainly but not by clinical status. In other words, we found a non-random distribution of genomes along the core phylogenetic tree which was found to be associated with intrinsic (selection of specific virulence genes) and extrinsic (environmental, urban vs. rural conditions) factors characterizing each genome. Our results also revealed a slightly greater virulence gene content in urban vs. rural areas, driven by isolates assigned to DAEC and EPECa. Despite the high diversity observed among our isolates in general, we identified several DAEC, ETEC and EIEC/Shigella clonal complexes that were significantly ($p<0.05$) or exclusively associated with diarrhea relative to control samples and encoded the hallmark virulence factors (VF) of the pathogroup, suggesting that they might have caused small-scale outbreaks in both rural and urban settings.

Our collection of genomes obtained from individuals with diarrhea was significantly populated with DAEC strains. In addition, the epidemiological component of the EcoZUR study showed that DAEC was significantly associated with diarrhea in urban areas (unpublished data). Consistent with the latter finding, we found that DAEC strains circulating in urban regions (Quito and Esmeraldas) harbored in general more virulence genes than their rural counterparts and that they were most likely to be clustered together in specific clades within the phylogeny, usually forming clonal complexes. Collectively, these findings indicated that DAEC is actually an important pathotype causing disease in Ecuador, and probably other parts of South America. However, DAEC is usually overlooked as a pathogen because its main mechanism of pathogenicity does not involve the secretion of specific toxins. Accordingly, almost all information available for DAEC pathotype has come from experiments *in vitro* and not field studies like our study.

Despite finding no association between clinical status and phylogenetic groups at large scale, we did observe the occurrence of several highly related, clustering of

isolates in the core phylogeny for DAEC, ETEC and EIEC, which were mostly found in diarrhea cases, defined here as clonal complexes. These clonal complexes encoded the hallmark virulence factors (VF) of the pathogroup, which suggested that they might have caused small-scale outbreaks in both rural and urban settings. For example, the EcoZUR epidemiological data showed a large association of EIEC with diarrhea. EIEC isolates also formed a discrete clonal complex within the B1 clade in the core genome phylogeny. These observations were consistent with the findings from the GEMS study (13), where authors reported that *Shigella* was among the top five pathogens responsible for diarrheal diseases in sub-Saharan Africa and Southeast Asia.

Further, our study showed that DAEC ($p=0.04$) and EPECa ($p=0.009$) strains circulating in urban areas carry in general, more virulence genes than those strains circulating in rural areas. This finding requires further, more detailed investigation as it likely has important implications for epidemiology and public health. It has been predicted that the most rapid growth in urban populations will occur in developing countries in the next 10 years. Therefore, determining whether urban areas promotes the establishment of more virulent genotypes than rural regions is critical in order to evaluate potential risks and to improve city planning and/or surveillance programs that can help to decrease the burden of communicable disease. Smith et al. (unpublished data) also showed that travelling to urban settings in the same region of Ecuador was a risk factor for disease, suggesting that investments in improving conditions in urban areas could have cascade effects in rural areas as well.

4.6 Acknowledgments

We thank to all participants from Ecuador who agreed to participate in the EcoZUR project and donated their time and stool specimens for the successful completion of the study. We would like to thank the local authorities of Ministry of Public Health for providing us access to the facilities in their hospitals. We would also like to thank Denys Tenorio, Mauricio Ayoví, Xavier Sanchez, Edison Puebla, and Kate Bohnert for their assistance with carrying out the field portions of the study. Funding for this study was provided by National Institute for Allergy and Infectious Diseases grant 1K01AI103544 and Colciencias through a doctoral fellowship to

APG. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

.

4.7 References

1. Conway, T. and Cohen, P.S., 2015. Commensal and pathogenic Escherichia coli metabolism in the gut. *Microbiology spectrum*, *3*(3).
2. Rossi, E., Cimdins, A., Lüthje, P., Brauner, A., Sjöling, Å., Landini, P. and Römling, U., 2018. "It'sa gut feeling"–Escherichia coli biofilm formation in the gastrointestinal tract environment. *Critical reviews in microbiology*, *44*(1), pp.1-30.
3. Tenaillon, O., Skurnik, D., Picard, B. and Denamur, E., 2010. The population genetics of commensal Escherichia coli. *Nature Reviews Microbiology*, *8*(3), p.207.
4. Jandhyala, S.M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M. and Reddy, D.N., 2015. Role of the normal gut microbiota. *World journal of gastroenterology: WJG*, *21*(29), p.8787.
5. Hudault, S., Guignot, J. and Servin, A.L., 2001. Escherichia coli strains colonising the gastrointestinal tract protect germfree mice againstSalmonella typhimuriuminfection. *Gut*, *49*(1), pp.47-55.
6. Kaper, J.B., Nataro, J.P. and Mobley, H.L., 2004. Pathogenic escherichia coli. *Nature reviews microbiology*, *2*(2), p.123.
7. Fleckenstein, J.M., 2013. Enterotoxigenic Escherichia coli. In *Escherichia coli (Second Edition)* (pp. 183-213).
8. Pasqua, M., Michelacci, V., Di Martino, M.L., Tozzoli, R., Grossi, M., Colonna, B., Morabito, S. and Prosseda, G., 2017. The intriguing evolutionary journey of enteroinvasive E. coli (EIEC) toward pathogenicity. *Frontiers in microbiology*, *8*, p.2390.
9. Kaur, P., Chakraborti, A. and Asea, A., 2010. Enteroaggregative Escherichia coli: an emerging enteric food borne pathogen. *Interdisciplinary perspectives on infectious diseases*, *2010*.
10. Pearson, J.S., Giogha, C., Wong Fok Lung, T. and Hartland, E.L., 2016. The genetics of enteropathogenic Escherichia coli virulence. *Annual review of genetics*, *50*, pp.493-513.
11. Servin, A.L., 2014. Pathogenesis of human diffusely adhering Escherichia coli expressing Afa/Dr adhesins (Afa/Dr DAEC): current insights and future challenges. *Clinical microbiology reviews*, *27*(4), pp.823-869.
12. Smith, J.L., Fratamico, P.M. and Gunther IV, N.W., 2014. Shiga toxin-producing Escherichia coli. In *Advances in applied microbiology* (Vol. 86, pp. 145-197). Academic Press.
13. Kotloff, K.L., Platts-Mills, J.A., Nasrin, D., Roose, A., Blackwelder, W.C. and Levine, M.M., 2017. Global burden of diarrheal diseases among children in developing countries: Incidence, etiology, and insights from new molecular diagnostic techniques. *Vaccine*, *35*(49), pp.6783-6789.
14. Lanata, C.F., Fischer-Walker, C.L., Olascoaga, A.C., Torres, C.X., Aryee, M.J. and Black, R.E., 2013. Global causes of diarrheal disease mortality in children< 5 years of age: a systematic review. *PloS one*, *8*(9), p.e72788.

15. Riveros, M., García, W., García, C., Durand, D., Mercado, E., Ruiz, J. and Ochoa, T.J., 2017. Molecular and Phenotypic Characterization of Diarrheagenic Escherichia coli Strains Isolated from Bacteremic Children. *The American journal of tropical medicine and hygiene*, *97*(5), pp.1329-1336.
16. Acosta, G.J., Vigo, N.I., Durand, D., Riveros, M., Arango, S., Zambruni, M. and Ochoa, T.J., 2016. Diarrheagenic Escherichia coli: prevalence and pathotype distribution in children from peruvian rural communities. *The American journal of tropical medicine and hygiene*, *95*(3), pp.574-579.
17. Torres, A.G., 2017. Escherichia coli diseases in Latin America—a 'One Health'multidisciplinary approach. *Pathogens and disease*, *75*(2).
18. Eisenberg, J.N., Cevallos, W., Ponce, K., Levy, K., Bates, S.J., Scott, J.C., Hubbard, A., Vieira, N., Endara, P., Espinel, M. and Trueba, G., 2006. Environmental change and infectious disease: how new roads affect the transmission of diarrheal pathogens in rural Ecuador. *Proceedings of the National Academy of Sciences*, *103*(51), pp.19460-19465.
19. Neiderud, C.J., 2015. How urbanization affects the epidemiology of emerging infectious diseases. *Infection ecology & epidemiology*, *5*(1), p.27060.
20. Zlotnik, H., 2017. World urbanization: trends and prospects. In *New Forms of Urbanization* (pp. 43-64). Routledge.
21. UN, 2014. *World Urbanization Prospects: The 2014 Revision-Highlights*. UN.
22. Enserink, R., Scholts, R., Bruijning-Verhagen, P., Duizer, E., Vennema, H., de Boer, R., Kortbeek, T., Roelfsema, J., Smit, H., Kooistra-Smid, M. and van Pelt, W., 2014. High detection rates of enteropathogens in asymptomatic children attending day care. *PLoS One*, *9*(2), p.e89496.
23. Praharaj, I., Revathy, R., Bandyopadhyay, R., Benny, B., KO, M.A., Liu, J., Houpt, E.R. and Kang, G., 2018. Enteropathogens and Gut Inflammation in Asymptomatic Infants and Children in Different Environments in Southern India. *The American journal of tropical medicine and hygiene*, *98*(2), pp.576-580.
24. Buffalo, V., 2014. Scythe-A Bayesian adapter trimmer (version 0.994 BETA).
25. Cox, M.P., Peterson, D.A. and Biggs, P.J., 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC bioinformatics*, *11*(1), p.485.
26. Peng, Y., Leung, H.C., Yiu, S.M. and Chin, F.Y., 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), pp.1420-1428.
27. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. and Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, pp.gr-186072.
28. Zhu, W., Lomsadze, A. and Borodovsky, M., 2010. Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, *38*(12), pp.e132-e132.
29. Luo, C., Rodriguez-r, L.M. and Konstantinidis, K.T., 2014. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic acids research*, *42*(8), pp.e73-e73.
30. Rodriguez-R, L.M., Gunturu, S., Harvey, W.T., Rosselló-Mora, R., Tiedje, J.M., Cole, J.R. and Konstantinidis, K.T., 2018. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. *Nucleic acids research*.
31. Clermont, O., Gordon, D. and Denamur, E., 2015. Guide to the various phylogenetic classification schemes for Escherichia coli and the correspondence among schemes. *Microbiology*, *161*(5), pp.980-988.

32. Chen, Y., Ye, W., Zhang, Y. and Xu, Y., 2015. High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic acids research*, *43*(16), pp.7762-7768.
33. Rodriguez-R, L.M. and Konstantinidis, K.T., 2016. *The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes* (No. e1900v1). PeerJ Preprints.
34. Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), pp.1792-1797.
35. Price, M.N., Dehal, P.S. and Arkin, A.P., 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, *5*(3), p.e9490.
36. Letunic, I. and Bork, P., 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research*, *44*(W1), pp.W242-W245.
37. Chen, L., Xiong, Z., Sun, L., Yang, J. and Jin, Q., 2011. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic acids research*, *40*(D1), pp.D641-D645.
38. Jia, B., Raphenya, A.R., Alcock, B., Waglechner, N., Guo, P., Tsang, K.K., Lago, B.A., Dave, B.M., Pereira, S., Sharma, A.N. and Doshi, S., 2016. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic acids research*, p.gkw1004.
39. Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A. and Pevzner, P., 2016. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *bioRxiv*, p.048942.
40. Carattoli, A., Zankari, E., Garcìa-Fernandez, A., Larsen, M.V., Lund, O., Villa, L., Aarestrup, F.M. and Hasman, H., 2014. PlasmidFinder and pMLST: in silico detection and typing of plasmids. *Antimicrobial Agents and Chemotherapy*, pp.AAC-02412.
41. Carattoli, A., 2009. Resistance plasmid families in Enterobacteriaceae. *Antimicrobial agents and chemotherapy*, *53*(6), pp.2227-2238.
42. Szczepanowski, R., Braun, S., Riedel, V., Schneiker, S., Krahn, I., Pühler, A. and Schlüter, A., 2005. The 120 592 bp IncF plasmid pRSB107 isolated from a sewage-treatment plant encodes nine different antibiotic-resistance determinants, two iron-acquisition systems and other putative virulence-associated functions. *Microbiology*, *151*(4), pp.1095-1111.
43. Matamoros, S., Hattem, J.M., Arcilla, M.S., Willemse, N., Melles, D.C., Penders, J., Vinh, T.N., Hoa, N., Jong, M.D. and Schultsz, C., 2017. Global phylogenetic analysis of Escherichia coli and plasmids carrying the mcr-1 gene indicates bacterial diversity but plasmid restriction. *Scientific Reports*, *7*(1), p.15364.
44. Chiluisa-Guacho, C., Escobar-Perez, J. and Dutra-Asensi, M., 2018. First Detection of the CTXM-15 Producing Escherichia coli O25-ST131 Pandemic Clone in Ecuador. *Pathogens*, *7*(2), p.42.
45. Pitout, J.D. and DeVinney, R., 2017. Escherichia coli ST131: a multidrug-resistant clone primed for global domination. *F1000Research*, *6*.
46. Escobar-Páramo, P., Grenet, K., Le Menac'h, A., Rode, L., Salgado, E., Amorin, C., Gouriou, S., Picard, B., Rahimy, M.C., Andremont, A. and Denamur, E., 2004. Large-scale population structure of human commensal Escherichia coli isolates. *Applied and environmental microbiology*, *70*(9), pp.5698-5700.
47. Pallecchi, L., Lucchetti, C., Bartoloni, A., Bartalesi, F., Mantella, A., Gamboa, H., Carattoli, A., Paradisi, F. and Rossolini, G.M., 2007. Population structure and resistance genes in antibiotic-resistant bacteria from a remote community with

minimal antibiotic exposure. *Antimicrobial agents and chemotherapy*, *51*(4), pp.1179-1184.
48. Escobar-Páramo, P., Le Menac'H, A., Le Gall, T., Amorin, C., Gouriou, S., Picard, B., Skurnik, D. and Denamur, E., 2006. Identification of forces shaping the commensal Escherichia coli genetic structure by comparing animal and human isolates. *Environmental microbiology*, *8*(11), pp.1975-1984.
49. Gordon, D.M., Stern, S.E. and Collignon, P.J., 2005. Influence of the age and sex of human hosts on the distribution of Escherichia coli ECOR groups and virulence traits. *Microbiology*, *151*(1), pp.15-23.
50. Czeczulin, J.R., Whittam, T.S., Henderson, I.R., Navarro-Garcia, F. and Nataro, J.P., 1999. Phylogenetic analysis of enteroaggregative and diffusely adherent Escherichia coli. *Infection and immunity*, *67*(6), pp.2692-2699.
51. Levine, M.M. and Robins-Browne, R.M., 2012. Factors that explain excretion of enteric pathogens by persons without diarrhea. *Clinical Infectious Diseases*, *55*(suppl_4), pp.S303-S311.
52. Boots, M. and Mealor, M., 2007. Local interactions select for lower pathogen infectivity. *Science*, *315*(5816), pp.1284-1286.
53. Cressler, C.E., McLEOD, D.V., Rozins, C., Van Den Hoogen, J. and Day, T., 2016. The adaptive evolution of virulence: a review of theoretical predictions and empirical tests. *Parasitology*, *143*(7), pp.915-930

CHAPTER 5

METAGENOMIC-BASED IDENTIFICATION OF THE ETIOLOGICAL AGENT OF INFECTIOUS DIARRHEA AND MICROBIOME SIGNATURES CAUSED BY DIFFERENT *E. COLI* PATHOTYPES

5.1 Summary

Escherichia coli infectious diarrhea is an important contributor to child morbidity and mortality worldwide. While important advances have been made in characterizing the infectious mechanisms of *E. coli* as an enteric pathogen, a basic understanding of the compositional and functional alterations that occur in the intestinal microbiome during active infections is still limited. In addition, it remains unknown whether different pathotypes produce similar or distinct alterations due to their characteristic mechanism of infection and virulence factors. Here, we analyzed 38 fecal samples taken from young children in Northern Ecuador suffering infectious diarrhea relative to 42 age-matched, healthy controls. *E. coli* isolates were recovered from all disease samples and were categorized in three pathotypes by PCR and whole-genome sequencing, indicating that they were the etiological agents based on conventional testing. The pathotypes were Diffusely Adherent (DAEC), Enteropathogenic (EPEC) and Enterotoxigenic (ETEC) *E. coli*. To validate these traditional culture-based methodologies, we obtained companion shotgun metagenomes and inspected for signatures of *E. coli* infection defined as: i) a

higher abundance of the pathogenic isolate relative to reference commensal strain or control samples, ii) high prevalence of virulence factors and/or toxins, including the PCR-identified marker gene for DAEC, ETEC and EPEC in the metagenome-assemble population genome (MAG) assigned to *E. coli*, iii) reduced intra-specific diversity (high clonality) of the infectious *E. coli* population relative to healthy individuals and, iv) significant association of the isolate with disease vs. control samples (epidemiology). We found that although the pathogenic *E. coli* was PCR-detected in all diarrhea samples, only about ~50% of the samples actually presented these metagenomic features consistent with *E. coli* infection. The analysis of the remaining samples was not conclusive, which suggested that likely other viral, bacterial or eukaryotic pathogens could have been the etiological agents. Based on the former selected samples, we found taxonomic, but not metabolic signatures discriminating the microbiomes associated with DAEC and ETEC infections. *Fusobacterium mortiferum* and *Campylobacter concisus* were significantly more abundant in ETEC infections while *Bifidobacterium longum* and *Alloprevotella tannerae* were more enriched in DAEC infections. Overall, these results showed that shotgun metagenomics can assist in the identification of the etiological agent of infectious diarrhea (diagnostics) and revealed taxonomic signatures on the gut microbiome related to infections by different *E. coli* pathotypes.

5.2 Introduction

Diarrheal diseases remain a major public health issue worldwide, especially in developing countries, where poor sanitary conditions and limited access to clean water exacerbate the burden (1, 2). *Escherichia coli* (*E. coli*) is a gut commensal of vertebrates, including humans (3). *E. coli* can nevertheless cause a broad range of diseases including intestinal and extra-intestinal infections. The relevance of this microorganism in diarrheal disease has been previously illustrated. For example, the Global Enteric Multicenter Study (GEMS) demonstrated that *Shigella* (a distinct lineage within the *E. coli* phylogenetic clade) and Enterotoxigenic *E. coli* (ETEC) producing heat stable toxin (ST) (either alone or in combination with heat labile toxin (LT)) are among the most important pathogens associated with moderate-to-severe diarrhea (MSD) in children younger than five years of age living in developing countries (4). Yet, little is known about the effect(s) of diarrheagenic *E. coli* on the gut microbiome such as how

the structure of the community changes during active infections, especially in young children.

While commensal *E. coli* is typically a minor component of the colonic gut microbiome in humans, where it represents less than 0.1% of the total bacterial cells (~10^8 cells/g) (5), during enteric infections with invading pathogenic *E. coli* strains or another enteric pathogen, the overall signal of *E. coli* in the gut microbiome can increase substantially (6, 7), allowing the detection and recovery of strain(s) that could accurately represent the invading population. In addition, quantifiable shifts in the proportion and diversity of the entire gut microbial community in response to the infection disturbance have been also observed (6-8). Therefore, characterizing alterations in the host gut microbiome in addition to the abundance, intra-population diversity and virulence content of the invading *E. coli* population during an intra-intestinal infection is important to better understand the overall dynamics of the gut microbial populations and for diagnostic testing.

Although most diarrheal cases self-resolve relatively quickly and hence, do not require typing of the causative agent(s), there are instances where acute diarrheal infections could lead to mortality, making detailed investigations of the causative agents and their signatures necessary. Such cases include foodborne outbreaks linked to contaminated food in developed countries and acute child diarrhea in the developing world. Diagnostic testing for enteric pathogens in these cases has relied for decades on culture-based techniques and this approach is the foundation for public health surveillance. It becomes important to realize that these approaches will frequently miss moderately or distantly related relatives of the reference culture. Further, and perhaps more importantly, pathogens, including *E. coli* (9), can quickly loose cultivability upon exposure to laboratory conditions, hampering the full characterization of infections (63). Accordingly, a total of 38.4 million cases of foodborne illness per year cannot be attributed to specific causes, and the proportion caused by yet-to-be-described microbial agents is unknown. Clinical metagenomics, the study of the genetic material recovered directly from a clinical specimen (21), has emerged as a powerful tool in clinical microbiology in the last few years, allowing researchers to explore the complex microbial communities of the human gastrointestinal tract bypassing the need for isolation or culturing (64). Its strength as potential diagnostic method relies in that it can not only

detect and recover the genome sequence of enteric pathogens from clinical specimens, but also allows profiling the surrounding microbial community, which could be also important for diagnosis and better understanding of the disease progression.

Based on virulence factors, pathogenic *E. coli* have been divided into six distinguishable pathotypes: Enterotoxigenic *E. coli* (ETEC), Enteroinvasive *E. coli* (EIEC), Enteroaggregative *E. coli* (EAEC), Enteropathogenic *E. coli* (EPEC), Diffusely adherent *E. coli* (DAEC) and Shiga-toxin producing *E. coli* (STEC) (6-12). From these, EPEC and ETEC have received special attention in developing countries as they are main pathogens causing diarrhea in children under five (1-4). While a variety of pathogenic *E. coli* can directly or indirectly cause diarrhea, i.e., the production of net water secretion towards the lumen, via different physiological mechanisms, it remains undetermined whether different *E. coli* pathotypes produce similar or perhaps, distinct alterations in the indigenous microbiome due to their characteristic mechanism of infection and virulence. Much of the work to date on the gut microbiome has focused on the relationship of the gut community composition to chronic diseases or conditions such as obesity (13, 14), malnutrition (15), or inflammatory bowel disease (IBD) (16-18). A much more limited number of studies have examined the impacts of infectious diarrhea on the gut microbial communities of young children (19, 20) or the existence (or not) of pathogen-specific signatures of the disturbed gut environment.

During the interaction with the intestinal epithelial cells, in the phase of attachment and colonization, distinctive machineries of pathogenicity are known to be used, depending on whether the pathogen invades the cell, produces biofilms or secrete toxins. For example, Enterotoxigenic *E. coli* (*ETEC*) adheres to the small bowel mucosa and delivers secretory enterotoxins (7). Enterohemorrhagic *E. coli* (*EHEC*) adheres intimately to the colonic mucosa and transduces a signal, resulting in secretory diarrhea. Concurrently, the organism releases Shiga toxin, resulting in local and systemic effects (12). Enteroaggregative *E. coli* (*EAEC*) adheres in a thick mucous gel (biofilm) and causes intestinal secretion and damage (9). Diffusely adherent *E. coli* (*DAEC*) has been shown to elicit elongation of microvilli *in vitro*, although this effect has not been demonstrated *in vivo*, and is considered, in general, not as virulent as other pathotypes (11). Enteropathogenic *E. coli* (*EPEC*) elicits the attaching and effacing lesion in the small bowel, resulting in intestinal secretion (10) and Enteroinvasive *E. coli* (*EIEC*)

invades the colonic mucosa, giving rise to inflammatory enteritis (8). These biological differences might disturb the gut microbial community in different, distinguishable ways. We hypothesized that the distinctive infectious mechanisms used by different *E. coli* pathotypes produce diagnostic taxonomic and/or functional genetic signatures in the sick gut microbiome that might inform about the pathotype of *E. coli* causing the infection.

As part of a large epidemiological, case/control study of diarrhea carried out over a period of 17 months in Northern coastal Ecuador (named EcoZUR for *E. coli* en Zonas Urbanas y Rurales), we examined the gut microbiome signatures of young children during diarrheal disease with infection by any of three major *E. coli* pathotypes, i.e., diffusely adherent DAEC, enterotoxigenic ETEC, and enteropathogenic EPEC and a group of age-matched control children. The main objectives of our study were to: 1) describe and compare the overall gut microbiome diversity between cases of diarrhea and controls using both 16S rRNA marker genes and whole shotgun metagenomes data, 2) determine cases of diarrhea where *E. coli* was most likely the causative agent of the disease based on a combination of metagenomics, isolate genome sequencing and epidemiological data, and 3) determine whether pathogen-specific signatures in the disease gut microbiome exist that distinguish among DAEC, EPEC and ETEC infections.

5.3 Methods and Materials

5.3.1 The EcoZUR study design

The EcoZUR study, which took place in Northern Ecuador over a period of 17 months, originally aimed to evaluate the risk of traveling between urban and rural settings in acquiring diarrheal infections and determining the distribution of *E. coli* pathotype in these settings. Fecal samples from individuals diagnosed with diarrhea and controls living in four regions along a gradient of urban to rural areas were taken and the traveling patterns of these individuals were catalogued. The regions included: Quito (Ecuador's capital) with approximately 1.6 million inhabitants; Esmeraldas, a coastal city in the northwest of Ecuador with 162,000 inhabitants; Borbón, a town in the Esmeraldas Province with ~7,000 inhabitants, and several villages (~150 villages) located along three main rivers: The Cayapas, the Santiago, and the Onzole, with 50 to 500

inhabitants each one. Participants were recruited between April 2014-September 2015 from the Ecuadorian Ministry of Health hospitals and/or clinics at each location. Individuals were recruited if they presented diarrhea, defined as three or more loose stools in a 24-hour period, and control subjects visiting the same medical facility with a non-diarrheal illness. Individuals were enrolled if they had no record of vomiting or antibiotic intake in the previous four weeks of the sample day. Data on demographics, medical history, water, sanitation, hygiene practices, animal contacts, and recent travel history information were collected from all participants. Surveys were carried out using Android devices and the Open Data Kit program (http://opendatakit.org). Prior to enrollment, all participants signed a consent document approved by the Institutional Review Boards of Emory University (IRB00065781) and Universidad San Francisco de Quito (USFQ) (2013-145M). The research protocol was also approved by the Ecuadorian Ministry of Health (MSP-DIS-2014-0055-O).

In the EcoZUR study, a total of 980 stool samples were taken. All diarrhea samples that resulted in PCR positive signal for the presence of any marker gene characterizing DAEC, ETEC and EPEC pathotypes and that were obtained from young children under six years old, were selected and prioritized for further analysis. A subset of 80 samples, which included diarrheal and control samples randomly chosen from the samples that met the above criteria of age and (for diarrhea samples) *E. coli* pathotype presence, were taxonomically screened by amplicon sequencing of the 16S rRNA gene. In addition, all diarrhea samples (n=38) and a subset of control samples (n=23) among the 80 samples were subjected to whole shotgun metagenomic sequencing for higher resolution (Appendix C.1). The diarrheal samples included 16 samples that were PCR-positive for *afa*, the virulence marker of DAEC (Diffusely Adherent *E. coli*), 10 samples positive for *bfp*, the marker gene for typical EPEC (Enteropathogenic *E. coli*) and 12 samples positive for *eltA* and/or *sta*, marker genes for ETEC (Enterotoxigenic *E. coli*). Children included in control group (n=42) were all diarrhea-free and PCR-negative for any of the pathotype markers characterized in the EcoZUR study. All available *E. coli* isolates cultured from stool samples from the selected cases of diarrhea were also sequenced for genomic characterization and comparison (Appendix C.2).

5.3.2 *E. coli* pathotype determination and Rotavirus detection

Fresh stool samples collected at the different locations in Northern Ecuador were incubated in *E. coli*-specific media and tested for different pathotypes based on the presence of specific virulence genes determined by PCR. For each stool sample, five lactose-positives colonies were isolated on MacConkey's agar media (MKL) and non-lactose fermenting isolates were further cultured and tested on Chromocult agar media (Merck, Darmsladt, Germany) (CC) for β-glucoronidase (MUG) activity. Colonies unable to ferment lactose were identified by biochemical tests as *Shigella* or *E. coli* using the API 20E test (BioMérieux, Marcy l'Etoile, France). The five colonies were pooled, re-suspended in 300µl of sterile distilled water, boiled for 10 minutes to release the DNA, and the resulting supernatant was used for PCR testing. Singleplex PCR assays were used on a set of nine different primers to detect the presence of virulence genes associated with each *E. coli* pathotype (Table 5.1). In addition, fresh stool samples were also tested for rotavirus antigens using the RIDA Quick Rotavirus test (r-biopharm, Darmstadt, Germany). Positive pools for *eaeA* were subsequently tested for *stx1* and *stx2* genes for the differentiation of potential EHEC infections. If a pooled sample tested positive for any virulence factor, then each of the five isolates were re-tested individually to identify the specific isolate carrying the virulence gene.

5.3.3 DNA extraction, library preparation and sequencing

DNA from *E. coli* isolates was extracted using the Wizard Genomic DNA Purification kit (Promega). DNA for stool metagenomes was extracted from a homogenized stool mix using the MoBIO PowerSoil DNA isolation Kit. In both cases, the purity and concentration of the DNA was estimated using a NanoDrop spectrophotometer (Thermo Scientific) and the Qubit 2.0 dsDNA high-sensitivity assay (Invitrogen, Carlsbad, CA). Libraries for 16S rRNA gene amplicon sequencing were amplified and sequenced using 16S rRNA primers 515F (5′-GTGCCAGCMGCCGCGGTAA-3′) and 806R (5′-GGACTACHVGGGTWTCTAAT-3′), which target the V4 region of the gene (~254 bps) and are tailed with Illumina adapters P5 and P7, index, pad and linker sequences as previously described in (26).

Table 5.1. Genes and primers used in the PCR identification of *E. coli* pathotypes

E. coli	Gene	Primer sequence 5 – 3'	Size (bp)	Reference
EAEC	*aggR*[a]	5'GTATACACAAAAGAAGGAAGC3' 5'ACAGAATCGTCAGCATCAGC3'	254	22
ETEC	*eltA*[b]	5'GCGACAAATTATACCGTGCT3' 5'CCGAATTCTGTTATATATGT3'	708	23
	sta[c]	5'CTGTATTGTCTTTTTCACCT3' 5'GCACCCGGTACAAGCAGGAT3'	182	24
EPEC	*bfp*[d]	5'CAATGGTGCTTGCGCTTGCT3' 5'GCCGCTTTATCCAACCTGGT3'	324	24
	eaeA[e]	5'GACCCGGCACAAGCATAAGC3' 5'CCACCTGCAGCAACAAGAGG3'	384	24
EIEC	*ipaH*[f]	5'GCTGGAAAAACTCAGTGCCT3' 5'CCAGTCCGTAAATTCATTCT3'	424	22
DAEC	*afa*[g]	5'GCTGGGCAGCAAACTGATAACTCTC3' 5'CATCAAGCTCTTTGTTCGTCCGCCG3'	750	25
STEC	*stx1*[h]	5'ATAAATCGCCATTCGTTGACTAC3' 5'AGAACGCCCACTGAGATCATC3'	180	24
	stx2[h]	5'GGCACTGTCTGAAACTGCTCC3' 5'TCGCCAGTTATCTGACATTCTG3'	255	24

[a]Transcriptional activator of aggregative adherence fimbria I; [b]Heat-labile enterotoxin A; [c]Heat-stable enterotoxin; [d]Bundle-forming Pili; [e]Gamma intimin; [f]Invasion Plasmid Antigen; [g]Afimbrial adhesin; [h]Shiga-Toxin.

PCR amplifications were performed in duplicates to a final volume of 20 µl containing 0.5U of AccuPrime-Pfx polymerase, 1X AccuPrime reaction mix, 200 nM of each primer and 1 µl of template DNA. Amplification conditions included an initial denaturation of 2 min at 95°C, followed by 25 cycles of 95°C for 20s, 55°C for 30s and 72°C for 30s, and a single final extension step at 72°C for 6 min. Specific amplification was verified by agarose gel electrophoresis and duplicate samples pooled and purified using the Sequalprep Normalization plate kit (ThermoFisher Scientific, USA) according to manufacturer's instructions. The concentration of purified DNA was measured using a Qubit fluorometer (Thermo Fisher Scientific, USA). Purified amplicons were then pooled in equimolar concentration (2 nM) and 10 pM containing 5% PhiX control DNA was sequenced on an in-house Illumina MiSeq instrument running the MiSeq control

software v2.4.0.4 (MCS) and using a MiSeq reagent v2 kit for 500 cycles (2 x 250 bp paired end run) (Illumina Inc., San Diego, CA). Adapter trimming and demultiplexing of sequenced samples were carried out by the MCS.

For isolates and metagenome DNA sequencing, libraries were prepared using the Illumina Nextera XT DNA library preparation kit (Illumina) according to manufacturer's instruction. After this, libraries were run on a high sensitivity DNA chip using the Bioanalyzer 2100 instrument (Agilent) to determine library insert sizes. An equimolar mixture of the isolates libraries (final loading concentration of 10 pM) was sequenced on an Illumina MiSeq instrument (School of Biological Sciences, Georgia Institute of Technology), using a MiSeq reagent v2 kit for 500 cycles (2 x 250 bp paired end run). Metagenomic libraries were sequenced in the Illumina HiSeq 2500 instrument in the rapid run mode for 300 cycles (150 bps, paired-end mode).

5.3.4 Read quality control, assembly and taxonomic annotation

Raw fastQ reads from both *E. coli* isolates and stool metagenomes were quality-trimmed using the SolexaQA++ package (27). Specifically, the scripts *dynamicTrim.pl* and *LengthSort.pl* were used to trim individual reads to the longest continuous segment for which phred quality score was greater or equal to 20 (Q≥20, which represents ~99% accuracy per nucleotide position). Reads shorter than 50 bps were discarded.

Both isolate and metagenome libraries were processed using MiGA (Microbial Genomes Atlas), a recently developed system for data management and processing of microbial genomes and metagenomes (http://microbial-genomes.org/) (28). In MiGA, quality-filtered reads are *de novo* assembled using IDBA-UD with pre-corrections (29) and protein-coding sequences predicted using MetaGeneMark (30). In addition, MiGA reports the percent of contamination and completeness for genome sequences based on recovery of lineage-specific marker genes. 16S rRNA gene sequences are also identified using barrnap 0.6 (https://github.com/tseemann/barrnap) and classified using the RDP classifier (33). In addition, through MiGA all predicted genes from the assemblies are taxonomically annotated using MyTaxa (31) and the taxonomic affiliation of adjacent genes (in windows of 10 genes) in the genome assembly are estimated, which allows the user to manually inspect the genomes for possible contamination

through the so called '*MyTaxaPlots'* barplots.

5.3.5 16S rRNA library processing

16S rRNA libraries were processed using the default parameters in the Quantitative Insights into Microbial Ecology (Qiime2) pipeline (*https://qiime2.org/).* In brief, 16S rRNA gene sequencing data was first denoised and quality filtered using Dada2, an algorithm that uses a statistical model for correcting errors introduced in the Illumina amplicon sequencing and infers underlying sample sequences clustering highly similar sequence variants (SV) (32). After quality control, a feature table of sequence variants that are 100% identical was generated. The naïve Bayesian RDP classifier (33), trained with the GreenGenes database (gg_13_8) at a 50% confidence, was used to assign taxonomy to the generated cluster of sequence variants (SVs). Low abundance SVs (those SVs with a total count fraction lower than 0.005% of the total) were removed from the final feature table as suggested by (34). Alpha-diversity estimates were computed in rarefied libraries to avoid bias in the detection of differentially abundant taxa given the underlying differences in library size. SV richness was estimated using Faiths' phylogenetic diversity (35). Beta-diversity analyses were performed using different distance metrics, including Bray-Curtis, Jaccard, Weighted and Unweighted UNIFRAC. Resulting distance matrices were visualized through principal coordinate analysis plots (PCoA). The same distance matrices were then used to conduct a statistical test of dissimilarities, i.e., a permutational multivariate non-parametric analysis of dissimilarities PERMANOVA (36), performed using the R package Vegan (37). Differentially abundant OTUs (or sequence variants) between diarrhea and control samples were identified using STAMP, a data analysis metagenomic software that reports differentially abundant features using effect sizes and confidence intervals (38).

5.3.6 Population genome binning

MaxBin2 (39) was used to bin previously assembled contigs into metagenome-assembled genomes (MAGs) for the recovery of *E. coli* population genome with a minimum contig length threshold of 2,000 bps. Prior to binning, Bowtie 2 was used to align short-read sequences to assembled contigs (40) and SAMtools was used to sort and convert SAM files to BAM format (41). Sorted BAM files were then used to calculate the coverage (mean representation) of each contig in each sample metagenome. The

quality of each resulting MAG was evaluated with CheckM v1.0.3 (42) using taxonomy-specific workflow for '*Escherichia coli*'. Only *E. coli* MAGs with a higher quality score than 60 (from Parks et al., 2017 (43); calculated as the estimated completeness minus five times the estimated contamination) were retained. The taxonomic affiliation of each MAG was then confirmed with MiGA, which uses a combination of the genome-aggregate Average Nucleotide Identity concept, or ANI, and the Average Amino-acid Identity, AAI, to taxonomically classify a query genomic sequence against its reference genome databases and find the closest match with *pval* <0.05. Protein coding genes on MAGs were predicted with Prodigal (44).

5.3.7 Microbial community composition and functional profile

Metagenome community taxonomic composition was assessed, in addition to the 16S rRNA gene level mentioned above, through clade-specific marker genes using MetaPhlAn2 v2.0 (45). Gene functional profiling of the microbial community was assessed using the pipeline implemented in HUMAnN2 (46), which implements searches against two main databases: UniRef, a comprehensive and non-redundant UniProt reference cluster for characterization of gene families (47), and MetaCyc, a highly curated database of experimentally elucidated metabolic pathways in all domains of life for evaluating the completeness of pathways (48). HUMAnN2 also implements MinPath, an algorithm for biological pathway reconstruction using protein family predictions (49). Metagenomic virulence profiling was examined using the Virulence Factors Database (VFDB, http://www.mgc.ac.cn/VFs/) (50) filtered for *E. coli* specifically. Metagenomic reads were mapped against the VF database and gene presence/absence was determined by the number of reads recruited by the VF genes (≥1X) and the length of the gene that was covered by reads (≥70%; lower gene abundance or coverage was considered gene absence).

5.3.8 Phylogenetic analysis of genomes of isolates and MAGs

Orthologous genes for isolates, MAGs and reference *E. coli* strains were identified using reciprocal best matches (RBM) with protocols detailed in (51). Sequences of core orthologous genes, i.e., genes present in all the genomes, were extracted and aligned using MUSCLE v3.8.35 (52). The resulting alignment was concatenated and trimmed

with Gblocks 0.91b (53) to remove noisy/uninformative regions. Phylogenetic reconstructions were estimated using FastTree version 2.1.7 (54) with 1,000 bootstrap replicates and the GTR-GAMMA substitution model for nucleotide sequences and visualized in iTOL (55).

5.4 Results and Discussion

5.4.1 Demographics of the individuals sampled

The demographic and clinical information for diarrhea and control groups are presented in Table 5.2. Individuals sampled from Quito and Esmeraldas composed the set of urban specimens while the ones sampled in Borbón and the villages composed the set of rural samples. The more urban centers (Quito and Esmeraldas) are densely populated regions with greater access to clean water, sanitation, roads, and medical infrastructure; whereas, the more rural regions (Borbón and rural villages) are less densely populated, with minimal sanitary infrastructure. Overall, no significant differences in demographic features including age, gender or race were found between cases and controls. However, and as expected, control samples were mostly retrieved from parents reporting improved household sanitation including flush toilets, personal latrines and/or septic systems (66.7%) relative to cases (39.5%). In addition, no significant differences in drinking water source or treatment were observed between cases and controls (Table 5.2).

Table 5.2. Characteristics of the study participants by site and disease status. No significant differences in demographic or socio-demographic characteristics were found between cases and controls within any given site. Total number of participants = 80; cases = 38 and controls = 42.

	Total n (%)	Cases n (%)	Controls n (%)
Demographics			
Region			
Urban (Quito + Esmeraldas)	38 (47.5)	20 (52.6)	18 (42.9)
Rural (Borbón + Villages)	42 (52.5)	18 (47.4)	24 (57.1)
Age			
0-24 months	36 (45.0)	22 (57.9)	14 (33.3)
25-60 months	28 (35.0)	12 (31.6)	16 (38.1)
61-180 months	16 (20.0)	4 (10.5)	12 (28.5)
Gender			
Male	42 (52.5)	16 (42.1)	26 (61.9)
Female	38 (47.5)	15 (39.4)	23 (60.5)
Race			
White	1 (1.2)	0 (0.0)	1 (2.4)
Black	29 (36.2)	12 (31.6)	17 (40.5)
Manaba	5 (6.2)	3 (7.9)	2 (4.8)
Indigenous	4 (5.0)	2 (5.3)	2 (4.8)
Mixed	41 (51.2)	21 (55.3)	20 (47.6)
Sociodemographics			
Family receives government welfare	15 (18.8)	1 (2.6)	14 (33.3)
Family member employed	26 (32.5)	12 (31.6)	14 (33.3)
Highest level household education			
Elementary	12 (15.0)	4 (10.5)	8 (19.0)
High School	45 (56.2)	24 (63.2)	21 (50.0)
University	23 (28.8)	10 (26.3)	13 (31.0)
Water and Sanitation			
Improved household sanitation[a]	43 (53.8)	15 (39.5)	28 (66.7)
Improved drinking water source[b]	72 (90.0)	36 (94.7)	36 (85.7)
Drinking water treatment[c]	28 (35.0)	15 (39.5)	13 (31.0)

[a]Improved = Flush toilets, personal latrine, and/or septic system. [b]Improved = Household tap, reclaimed rainwater, and/or purchased bottled water. [c]Options included boiling, chlorine usage, filtration, UV irradiation, larvicide treatment and/or settling techniques. However, 100% of those reported treating water used the boiling method.

5.4.2 High frequency of co-eluting human DNA in diarrhea samples

The metagenomes obtained showed differences in the proportion of microbial and human reads. The average percent of human reads detected in diarrhea group was 17.8% vs. 0.07% in the control group. The metagenomic recovery of large amount of human reads indicated infectious (as opposed to osmotic) diarrhea. Case samples with large fraction of human reads yielded mostly DAEC(+) and EPEC(+) pathotype isolates (Figure 5.1A; See also next section for DAEC infection). Next, we evaluated whether or not the fraction of human reads detected in each dataset correlated with the severity of the disease, measured as the number of days with diarrhea previous to the sampling day and the detection of blood and/or mucus in the specimen. Our results showed no significant correlation with the number of days with diarrhea or presence of blood. However, we did observe a significant increase in the fraction of human reads in samples with detection of mucus in stool specimens versus those with no mucus detected (Welch's two sample test, *pval=0.004*). This observation might be related to the intestinal niche that *E. coli* (and other enteric pathogens) colonizes. Previous studies based on fluorescent *in situ* hybridization microscopy (FISH) of thin sections of the cecum of streptomycin-treated mice have revealed that colonized *E. coli* do not attach to the epithelial cells, but growth takes place predominantly in the mucus layer instead (3, 56-58). Failure to penetrate the mucus and grow as dispersed cells prevented *E. coli* from effective colonization of the gut (57). The mucus layer itself is in a dynamic, yet highly regulated state, constantly being synthesized and secreted by specialized goblet cells, and it is degraded to a large extend by the indigenous intestinal microbes. Degraded mucus components are shed into the intestinal lumen and excreted in the feces (3, 58). Collectively, our observations suggested that during acute infectious diarrhea, a large fraction of mucus (presumably containing both epithelial and immune cells derived from tissue damage and/or pro-inflammatory responses) is washed out into the lumen and excreted in diarrhea possibly together with the pathogen.

Following removal of human reads, between 147Mb and 1.7Gb of reads per metagenome remained for analysis. To evaluate the fraction of the total extracted DNA from the stool sample that was sequenced (i.e., determine the coverage of the microbial community by sequencing), we estimated the sequencing coverage using Nonpareil, a read redundancy-based algorithm to estimate coverage (59). Although the sequencing coverage varied among samples and pathotype groups, generally ≥80% of the microbial

community was covered in the majority of samples, except for one EPEC sample that had very low coverage and was therefore, discarded from further analysis (Figure 5.1B). The coverage estimates also reflected differences in the complexity of the microbial communities, showing that DAEC group contained, in general, metagenomes with simpler microbial communities when compared to the other pathotypes, while control group comprised the individuals with the more complex (diverse) gut microbial communities. Overall, our results in community coverage suggested that our metagenomic sequencing effort was adequate to assess the microbial community disturbance during diarrhea and for potential pathogen detection and recovery.

5.4.3 Microbial community composition in diarrhea and control samples

To assess the overall microbial community composition and diversity in diarrhea and control groups, 16S rRNA gene amplicons sequences were analyzed. Comparison of the normalized relative abundance at the phylum level indicated that three major phyla dominated the gut microbial communities in both diarrhea and control individuals, namely *Bacteroidetes*, *Firmicutes* and *Proteobacteria*. In general, *Bacteroidia* and *Clostridia* were the two most abundant classes in all samples. However, the majority of samples in the diarrhea group also exhibited a large proportion of sequences belonging to *Gammaproteobacteria*, a bacterial group comprising several enteric pathogens, which suggested a possible underlying bacterial infection in individuals with diarrhea (Figure 5.2A).

Alpha-diversity analysis based on Faiths's phylogenetic diversity also revealed significant differences between case and control datasets. As expected, control stool samples presented, in general, a significantly higher microbial diversity than the cases (*Kruska-Wallis test, H=15.9, qval*= 6.06e-05) (Figure 5.2B). These results were consistent with the study reported by Pop and colleagues (2014) (19) where the gut microbiota composition in children from low-income countries in Sub-saharan Africa and Southeast Asia was assessed using 16S rRNA gene amplicon sequencing.

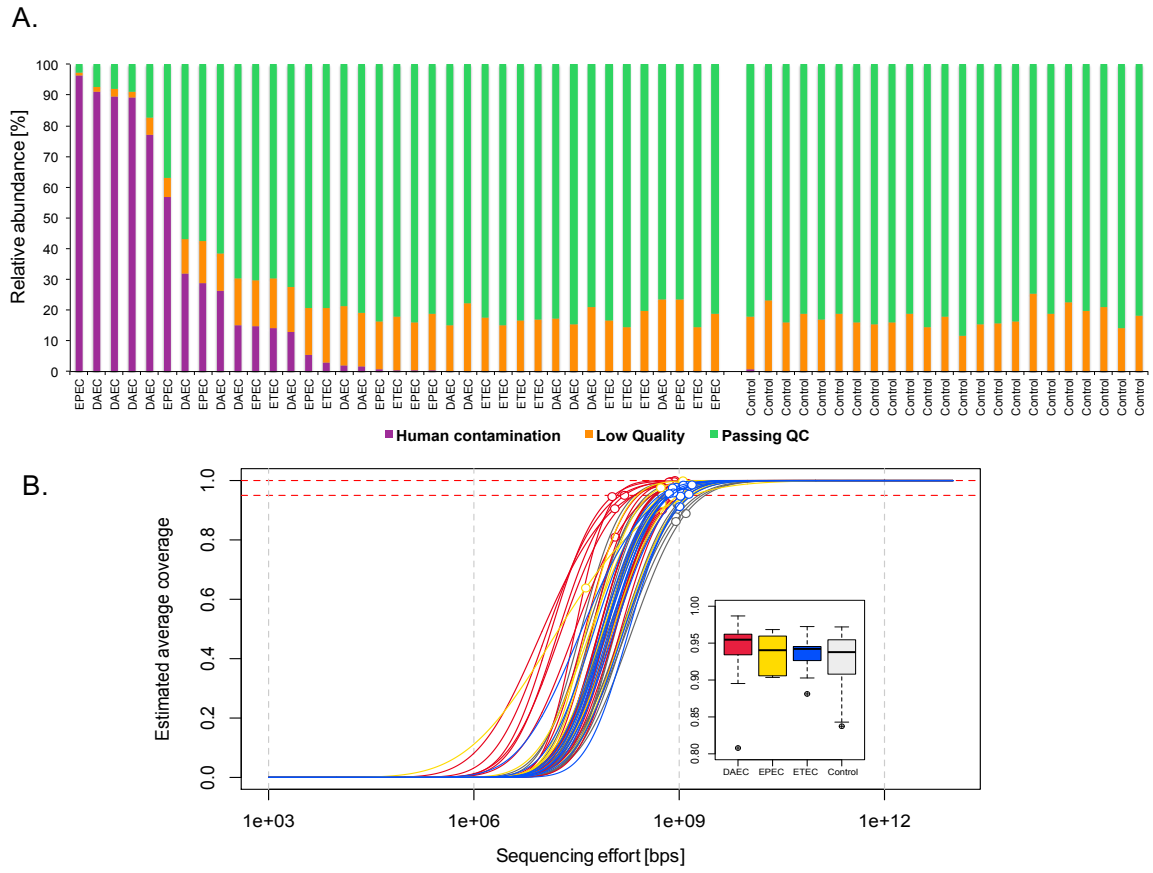


Figure 5.1. Abundance of human reads and estimated coverage of the metagenomic datasets obtained in this study. Panel (A) shows the assignment of recovered metagenomic raw reads to three groups: human (purple), discarded due to low quality (orange), and the fraction passing quality control and not be of human origin (green). Panel (B) shows the fitted Nonpareil curves and the estimated average coverage for each metagenome. The horizontal dashed lines indicate 100% (upper red line) and 95% (bottom red line) coverage. Empty circles indicate the size (x-axis) and estimated average coverage (y-axis) of the datasets, and the lines after that point are projections of the fitted model. Inset plot shows the distribution of estimated coverage values in a query subset of 1,000 reads per library in each pathotype and control groups. Note that samples were DAEC was isolated from showed less diverse communities when compared with other groups, including control samples.

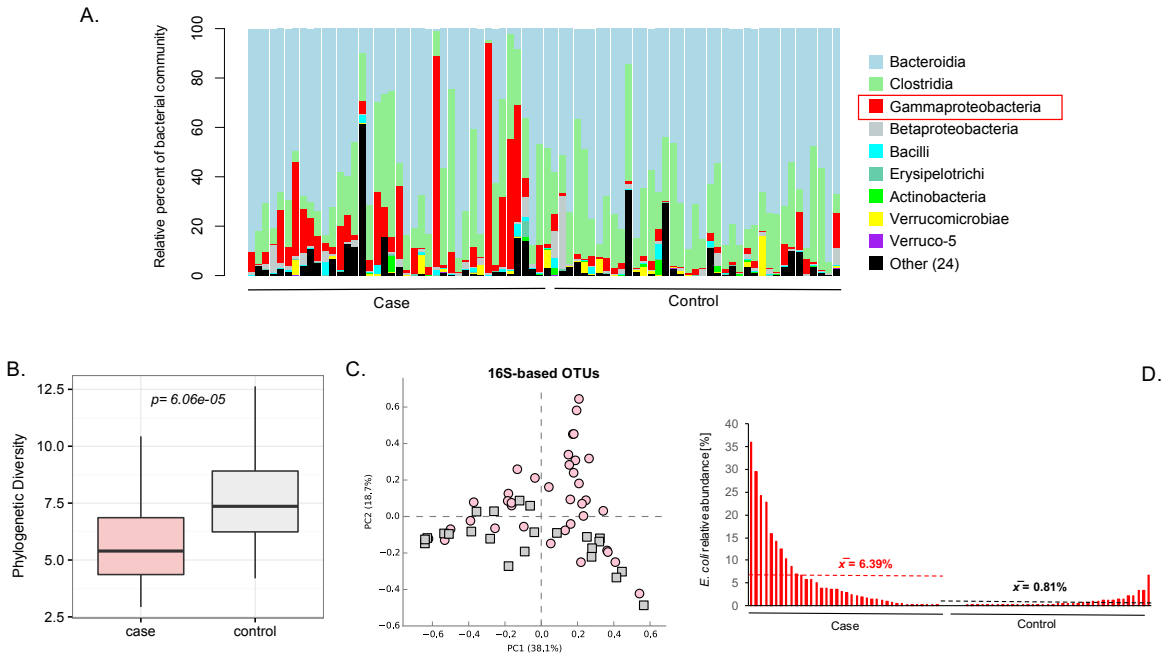


Figure 5.2. 16S rRNA gene-based microbial community composition differences between diarrhea and control samples. Panel (A) shows the relative abundance of bacterial groups classified at the class level for cases of diarrhea and control samples. Only the top ten more abundant phylogroups are displayed. Note that a higher abundance of *Gammaproteobacteria* was observed in case versus control groups. Panel (B) shows significant differences in the phylogenetic diversity between case and control samples. Consistent with previous literature, diarrhea samples presented lower community diversity than control ones. Panel (C) represents the overall community dissimilarity based on the taxonomic composition at the genus level. Pink circles represent cases of diarrhea and gray squares represent control samples. Panel (D) shows the estimated relative abundance percent of the 16S rRNA gene-based OTU (or sequence variant) taxonomically assigned to *E. coli*.

Analysis of the overall community dissimilarity based on the 16S rRNA gene-based taxonomic composition at the genus level also revealed significant differences between diarrhea and control groups (*PERMANOVA, pseudo-F=2.47, p=0.002*). Diarrhea samples clustered more closely among them than control samples, although

some overlap between several samples of the two groups was also observed (Figure 5.2C). Difference in the overall relative abundance of the 16S rRNA gene-based OTU taxonomically assigned to *Escherichia coli* in diarrhea and control samples were also detected. The estimated abundance of *E. coli* 16S rRNA gene reads in the diarrhea group was higher than the control group by about one order of magnitude (mean 6.39% of total reads in cases versus 0.81% in control, p= 0.009), and a few of the samples had as much as 30% of their totals sequences represented by the E. *coli* OTU. However, a large range in *E. coli* abundance in the diarrhea group was also observed, with several samples having very low *E. coli* abundances that were comparable to those in control samples (Figure 5.2D). Conversely, a couple of control samples had as much as 3-5% relative abundance of *E. coli*, which generated intriguing questions regarding the role of *E. coli* as the causative agent of disease. To follow up in these observations, we next aimed to identify samples where pathogenic *E. coli* was most likely the causative agent of the diarrheal disease.

5.4.4 Identification of disease samples where *E. coli* was the causative agent

Our observations of the overall community composition, diversity and estimated taxon abundance in the diarrheal samples (e.g., Figure 5.2) suggested that *E. coli* was likely not the causative agent of diarrhea disease in all (sick) individuals, even though *E. coli* was isolated from all case samples. In other words, isolation and PCR detection of *E. coli* pathotypes in stool samples from individuals with diarrhea might have reflected a stage of infection with *E. coli* but not necessarily disease caused by *E. coli*, or the recovery of a rare isolate *in-situ* that was not involved in infection. Therefore, we next aimed to determine the diarrheal cases that were attributable to pathogenic *E. coli* by assessing the metagenomic datasets for four main lines of evidence: first, the *in-situ* metagenomic abundance of the *E. coli* pathotype isolate based on the case metagenome that the isolate was recovered from should be higher than that in the control samples, after one accounts for reads representing commensal *E. coli* populations; the latter reads were identified by a competitive search against the isolate genome and that of the commensal *E. coli* strain HS (NC_009800.1). Second, the pathotype-specific toxins and virulence factors should be detectable in the metagenomic reads (no effect of assembly) at similar (or higher) abundances the corresponding isolate from the same sample and/or present in the *E. coli* MAG recovered from the metagenome. Third, the degree of intra-population diversity (or clonality) of the

pathogenic *E. coli* population should be lower (more clonal) compared to the *E. coli* population in control samples as it is often the case for infection-associated pathogens. The degree of clonality was estimated by calculating the average nucleotide identity of the metagenomic reads mapping to the reference genome (isolate or reference commensal) with percent identity between 90-100% (ANIr), i.e., reads representing the total *E. coli* population in a sample. Finally, the epidemiology of the clonal complex that the pathotype isolate was assigned to should be consistent with infection, i.e., other isolates in the same complex to be associated more strongly with disease than control samples (Appendix C.2, C.3 and C.4). Apperloo-Renkema et al (1990) reported that despite colonization resistance, humans are colonized on average with five different *E. coli* strains and there is a continuous succession of strains in individuals (61, revised in 3). Based on this finding, our working hypothesis was that control individuals carry in general, a more diverse population of *E. coli*, whereas diarrhea samples carry a more clonal pathogenic population that is present in higher abundance and encodes more virulence genes than the commensal population. We tested our hypothesis for three pathotype groups (DAEC, ETEC and EPEC) for which we obtained enough diarrheal samples (about a dozen samples/individuals per pathotype).

Diffusely adherent E. coli (DAEC) group

Our observations for the DAEC group showed consistent results with our working hypothesis for approximately 50% of the samples that DAEC isolates were obtained from (eight samples out of sixteen) (Figure 5.3). These samples corresponded to Q51, Q196, E158, Q56, Q65, E230, E124 and E27 and therefore, were selected as representative samples of DAEC infection. In general, this set of samples exhibited the following metagenomic signatures: 1) higher abundance of the pathogenic isolate compared to the reference commensal or the total *E. coli* population in the control samples (Figure 5.3E), on average (27.81% vs. 0.6%; Figure 5.3A); 2) recovery of high-quality *E. coli* MAGs that encoded the pathotype (*afaB-I*) and other *E. coli* (40 or more) virulence factors or the virulence factors were present in metagenomic contigs that were not binned into MAGs (Figure 5.3B); 3) reduced intra-population sequence diversity with ANIr values ≥ 99% for the isolate and usually lower values for the reference commensal genome (Figure 5.3C) and finally, 4) the recovered pathotype isolate(s) were generally

grouped in phylogenetic clusters where more isolates originated from cases of diarrhea than controls (Figure 5.3D, Appendix C.3).

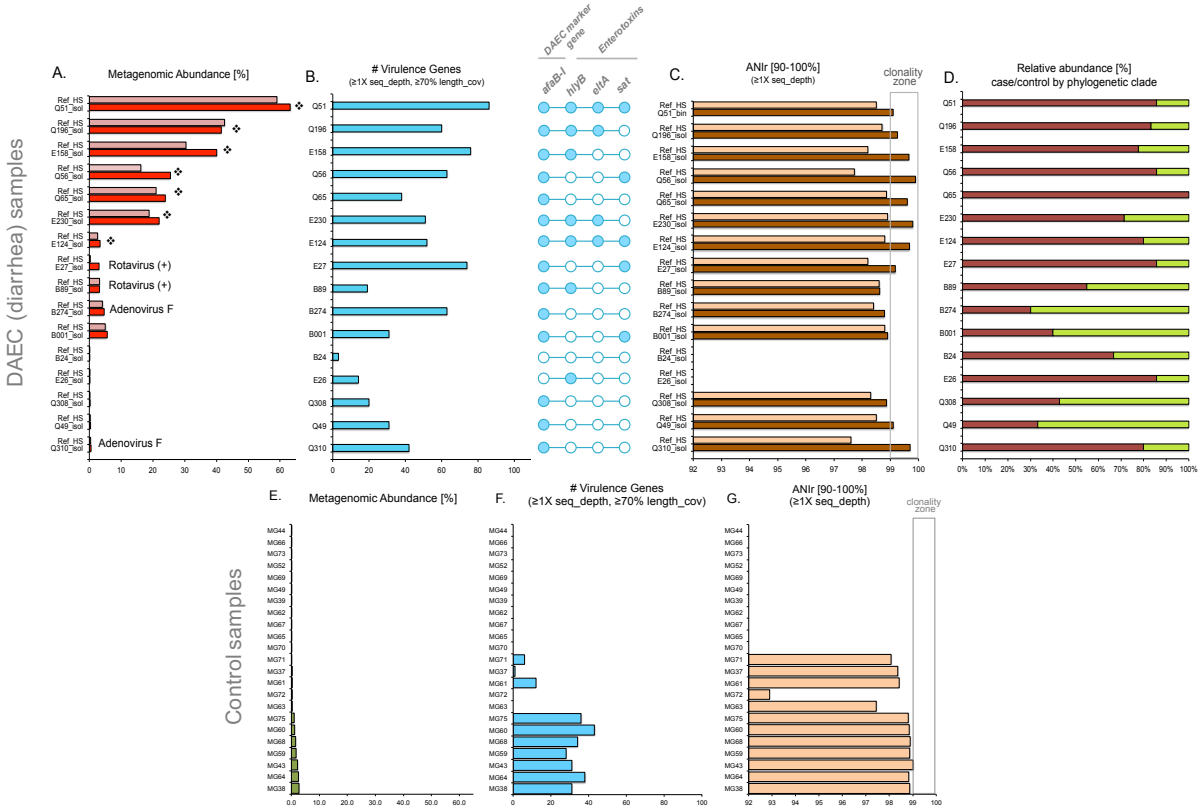


Figure 5.3. Characteristics of samples where DAEC was most likely the causative agent. Panel (A) shows the estimated metagenomic abundance of the reference commensal *E. coli* (strain HS, in light red) and the sample-specific DAEC isolate (in red) recovered from stool specimens, along with the Elisa-based detection of Rotavirus (+) and Adenovirus_F. Samples where high-quality *E. coli* MAGs were recovered are denoted by a star. Panel (B) shows the number of total *E. coli* virulence genes observed in the metagenome and an array of four hallmark virulence factors including the DAEC marker gene (*afaB-I*) and three enterotoxins, i.e., the hemolysin subunit B (*hlyB*), the heat-labile enterotoxin (*eltA*), and the secreted autotransporter toxin (*sat*). Panel (C) shows the estimated *E. coli* intra-population diversity measured by ANIr of reads against the reference commensal strain HS (light brown) and the isolate obtained from the

sample (dark brown). To avoid any potential bias by low *in-situ* abundance, only samples where the average sequence depth of the reference genome was ≥1X were evaluated for ANIr. Panel (D) shows the relative abundance of isolates from cases of diarrhea and control in the core-genome-based phylogenetic clade where the isolate was assigned to (epidemiology). Panels (E), (F) and (G**)** shows the same information than panel A, B and C for control samples.

Although some other DAEC-positive diarrhea samples showed ANIr values >99% and presence of the *afaB-I* gene (for example samples Q310 and Q49), they presented lower metagenomic abundances of the isolate compared with the control samples and/or other positive samples, and relatively lower number of virulence genes. In addition, for these low abundance populations no MAGs were recovered. Therefore, these samples were not conclusive with respect to whether or not the isolate was the etiological agent. Of particular interest was the observation that samples E27 and B89 were also positive for Rotavirus and samples B274 and Q310 showed a metagenomic signal of reads mapping to Adenovirus_F, a non-enveloped, double-stranded DNA virus causing acute gastroenteritis primarily in children (62). These results indicated that despite the PCR detection of DAEC marker gene in these individuals, other viral pathogens rather than *E. coli* might have been responsible for the diarrhea phenotype. Next, we evaluated whether any correlation existed between the percentage of human reads detected in the DAEC metagenomes and the estimated metagenomic abundance of the isolate, specifically in the set of eight samples that had strong evidence of DAEC-caused infection. Our results showed a significant, positive linear correlation between the two variables (*Pearson's R=0.83, p=0.01)* (Figure 5.4) suggesting that the fraction of human reads observed in the metagenome might be directly related with the infection by pathogenic DAEC strains.

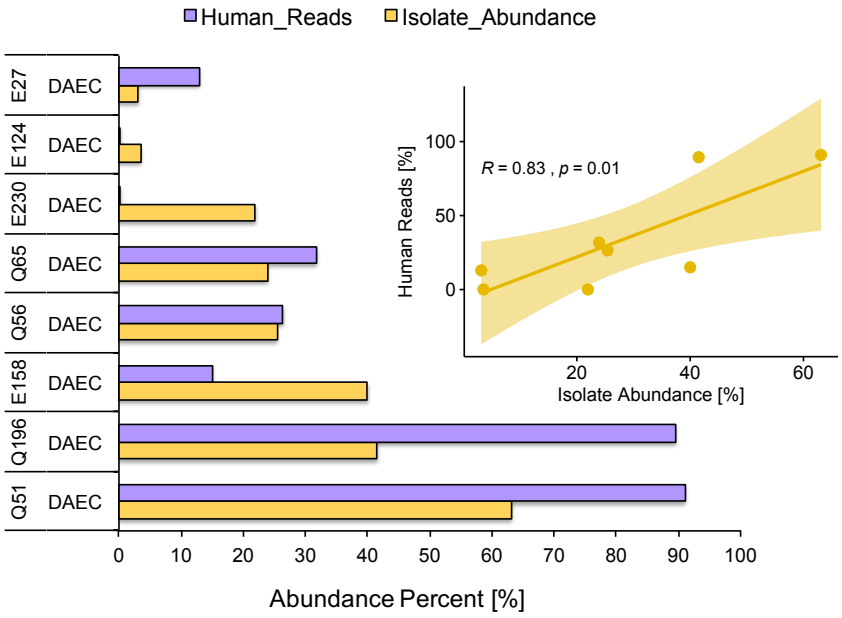


Figure 5.4. Correlation between recovered fraction of human metagenomic reads and DAEC pathogen abundance. The barplot shows the observed percent of the total metagenomic reads assigned to human (purple) and the estimated metagenomic abundance of the *E. coli* genome for the samples with strong evidence of DAEC infection. Inset plot shows the Pearson correlation analysis of the two variables, revealing a significant, positive linear correlation.

Enterotoxigenic E. coli (ETEC) group

A similar metagenomic assessment was performed in the ETEC pathogroup to identify representative samples of infection (Figure 5.5). In this sample set, the signal of *E. coli* infection was, in general, more clear than DAEC. Seven samples out of ten (70%) had strong evidence of infection caused by the ETEC isolate, i.e., samples E184, Q294, B295, B45, B62, B244 and B255. In these samples, at least one of the two ETEC marker genes, heat-labile (*eltA*) and/or heat-stable (sta) enterotoxins was detected, except for individual B200. The remaining three samples (i.e., B109, B68 and B200) showed very low abundance of *E. coli*, and/or the absence of at least one ETEC marker gene and/or relatively low clonality (ANIr <99%) and therefore, no strong evidence of *E. coli* infection. Of particular interest was the observation that in general, ETEC samples presented

lower average metagenomic abundance of the pathogen (5.07%) than DAEC samples (27,81%) by ~ 5 fold, on average, for the cases with clear evidence of *E. coli* pathotype infection. This observation suggested that ETEC infection might require a lower pathogen load in order to elicit disease than DAEC. In general, ETEC pathogens are widely recognized by the production of two enterotoxins: heat-labile (LT) and/or heat-stable (ST) enterotoxins (7). These proteins affect a variety of cellular processes that alter the concentration of important cellular messengers such as cyclic AMP, cyclic GMP and Ca^{2+} leading to increased ion secretion and acute watery diarrhea. DAEC, on the other hand, is characterized by the production of a fimbrial adhesin F1845 that is used to strongly attach to the cell surface. DAEC strains induce a cytopathic effect that is characterized by the development of long cellular extensions, which wrap around the adherent bacterial cell (11). These differences in mechanisms of pathogenicity might explain, at least in part, the differences in pathogen abundance that were observed in ETEC- and DAEC-dominated metagenomes.

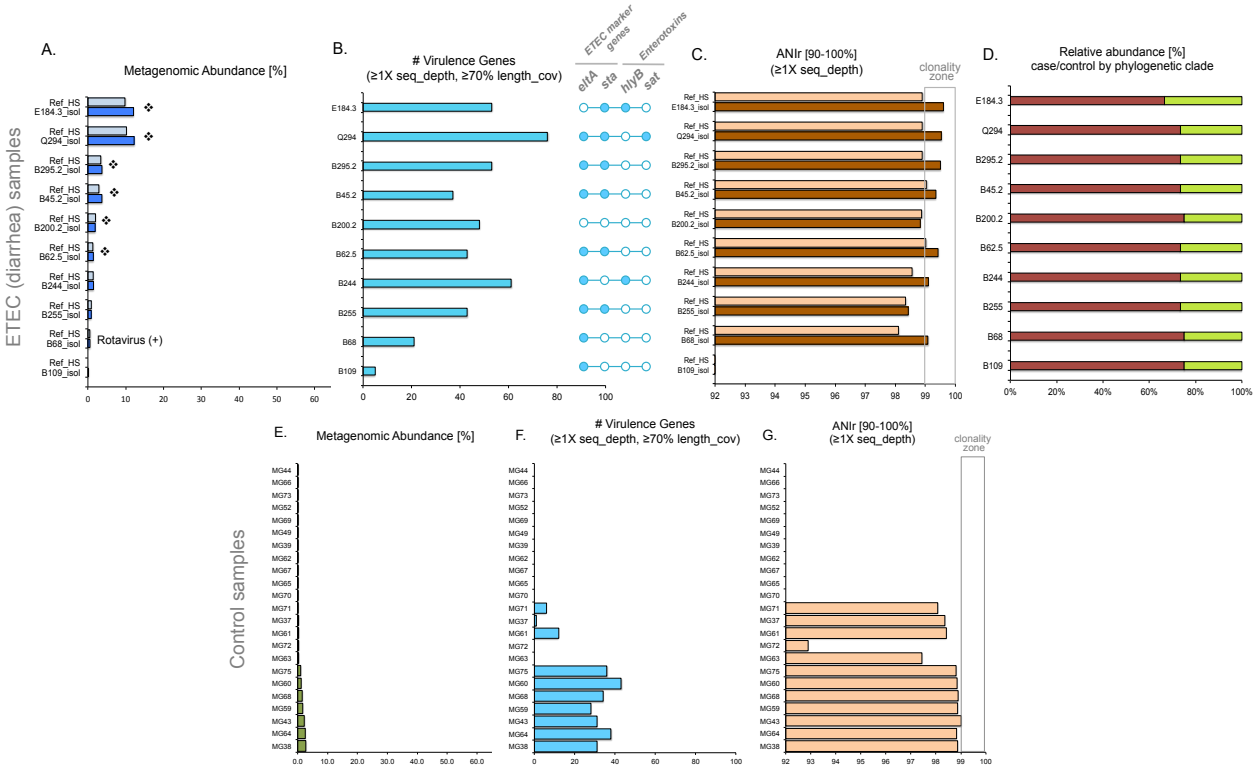


Figure 5.5. Characteristics of samples where Enterotoxigenic *E. coli* (ETEC) was most likely the causative agent. Panel (A) shows the estimated metagenomic abundance of the reference commensal *E. coli* (strain HS, in light blue) and the ETEC isolate (in blue) recovered from stool specimens, along with the Elisa-based detection of Rotavirus (+). Samples where high-quality *E. coli* MAGs were recovered are denoted by a star. Panel (B) shows the number of total *E. coli* virulence genes observed in the metagenome and an array of four hallmark virulence factors including the two ETEC marker genes (*eltA* and *sta*) and two additional enterotoxins (the hemolysin subunit B (*hlyB*) and the secreted autotransporter toxin (*sat*)). Panel (C) shows the estimated *E. coli* intra-population diversity measured by ANIr of reads against the reference commensal strain HS (light brown) and the isolate obtained from the sample (dark brown). To avoid any potential bias by low *in-situ* abundance, only samples where the average sequence depth of the reference genome was ≥1X were evaluated for ANIr. Panel (D) shows the relative abundance of isolates from cases of diarrhea and control in the core-genome-based phylogenetic clade where the sample-specific isolate was assigned to (Appendix C.4). Panels (E), (F) and (G) shows the same information than panel A, B and C for control samples.

Enteropathogenic E. coli (EPEC) group

The pathogenicity analysis performed in the EPEC pathogroup was the least conclusive (Figure 5.6). In general, the EPEC marker gene (*eaeA*) was recovered in only two samples (E187 and E162) out of eight in total (20%). However, even for the latter two sample, the analysis in intra-population diversity revealed no clonal population. Recovery of high-quality MAGs was possible only in one sample (R135) but less than 50% of the total number of virulence genes searched in the metagenomes were found in general. Given that no strong metagenomic signature of pathogenicity was observed in the majority of EPEC samples, we excluded this pathogroup from further analysis that focused on the detection of pathotype-specific gut microbiome signatures.

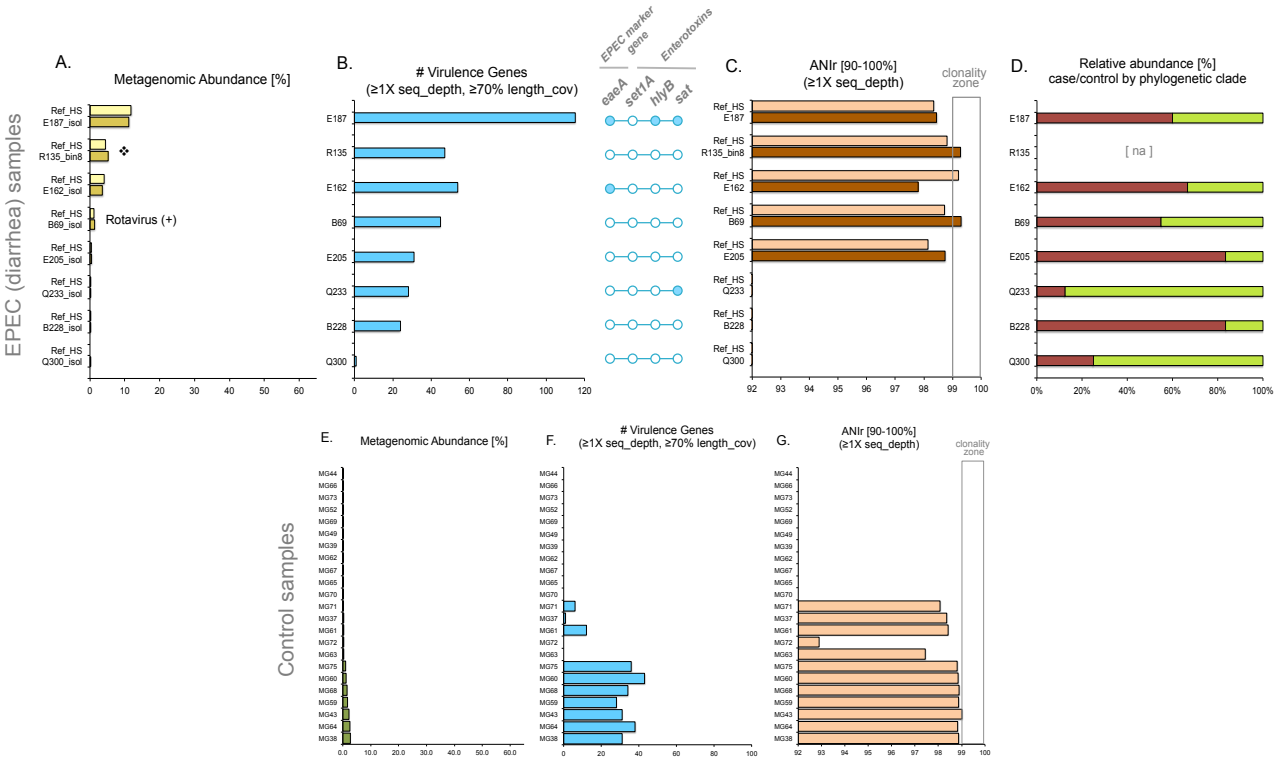


Figure 5.6. Characteristics of samples where Enteropathogenic *E. coli* (EPEC) was most likely the causative agent. Panel (A**)** shows the estimated metagenomic abundance of the reference commensal *E. coli* (strain HS, in light yellow) and the sample-specific EPEC isolate (in dark yellow) recovered from stool specimens, along with the Elisa-based detection of Rotavirus (+). Samples where high-quality *E. coli* MAGs were recovered are denoted by a star. Panel (B) shows the number of total *E. coli* virulence genes observed in the metagenome and an array of four hallmark virulence factors including the EPEC marker gene (*eaeA,* intimin) and three enterotoxins, i.e., the hemolysin subunit B (*hlyB*), the enterotoxin (*set1A*), and the secreted autotransporter toxin (*sat*). Panel (C) shows the estimated *E. coli* intra-population diversity measured by ANIr of reads against the reference commensal strain HS (light brown) and the obtained from the sample (dark brown). To avoid any potential bias by low *in-situ* abundance, only samples where the average sequence depth of the reference genome was ≥1X were evaluated for ANIr. Panel (D) shows the relative abundance of isolates from cases of diarrhea and control in the core-genome-based phylogenetic clade where the sample-

specific isolate was assigned to (Appendix C.5). Panels (E), (F) and (G) shows the same information than panel A, B and C for control samples.

5.4.5 Detection of differentially abundant phylogroups driving the differences between diarrhea and control samples

The identified samples with strong evidence of DAEC or ETEC infection were analyzed further in order to elucidate any signatures of the infection on the gut microbiome. First, we aimed to elucidate which taxa were contributing to the differences between diarrhea and control groups for this subset of samples. For this purpose, the metagenomes were analyzed for the presence of clade-specific marker genes, which allowed for a more accurate species-level characterization (see Materials and Methods). Differentially abundant features were reported if they had a corrected p-value ≤ 0.05 (q-val) and an effect size ≥0.8 based on the *Kruska-Wallis* test. In total, seven discriminative taxa between diarrhea and control samples were detected: *Prevotella copri (p=0.048), Escherichia coli (p=2.26e-5), Alloprevotella tannerae (p=0.046), Campylobacter concisus (p=0.011), Haemophylus haemolyticus (p=0.05), Fusobacterium mortiferum (p=0.025) and Bifidobacterium longum (p=0.040)* (Figure 5.7A). *Prevotella copri* was most strongly enriched in the control group while *E. coli* was the phylogroup mostly differentiating the diarrhea group, revealing an anti-correlation abundance pattern for these two species (*Speraman's rho= -0.39*, *p=0.012*). A principal component analysis of species abundance indicated that microbial communities from DAEC vs. ETEC (probable) infection clustered separately, in general (Figure 5.7B). Therefore, we next evaluted in more resolution the differences in the diarrheal metagenomes of the latter two groups after removing reads assigned to human and *E. coli*.

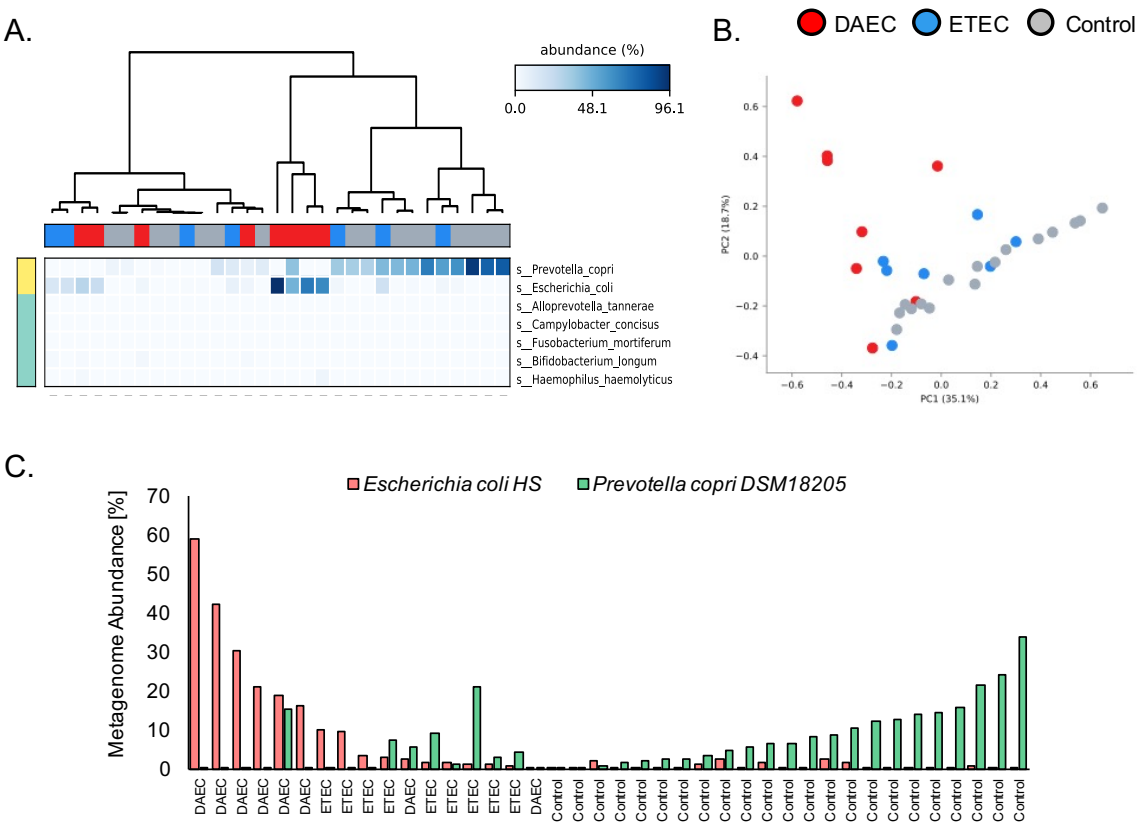


Figure 5.7. Differentially abundant taxa between *E. coli* infectious diarrheal and control samples. Panel (A) shows a heatmap with the relative abundance of the seven significantly differentially abundant species between diarrhea and control groups. Panel (B) shows a principal component plot of the taxonomic relatedness of samples with pathotype infection and controls using MetaPhlAn2. Panel (C) shows the estimated metagenomic abundance of *E coli*, using strain HS as the reference commensal genome to recruit *E. coli* reads (NC_009800.1) and *Prevotella copri*, using as reference the genome of strain *DSM18205*, a gut fermentative microbe (NZ_ACBX00000000.2), based on the number of metagenomic reads recruited by these genomes.

5.4.6 Lack of pathotype-specific signatures on gut microbiome distinguishing between DAEC and ETEC infections

Next, we aimed to determine whether pathotype-specific signatures in the gut microbiome existed that can distinguish among DAEC and ETEC infections. For this purpose, we profiled the taxonomic composition and metabolic pathways in the metagenomes of both diarrhea groups and compared with control samples after removing human and *E. coli* reads and normalizing for the size of the resulting metagenomic (sub-)datasets. The initial taxonomic characterization revealed at least four species that were discriminatory of DAEC infections versus ETEC ones (Figure 5.8). *Fusobacterium mortiferum* and *Campylobacter concisus* were significantly more abundant in ETEC infections (Figure 5.8AB) while *Bifidobacterium longum* and *Alloprevotella tannerae* were significantly more enriched in DAEC infections (Figure 5.8CD). Principal component analysis also revealed that ETEC metagenomes tended to be taxonomically more similar among them while DAEC samples showed more diversity. One DAEC sample (E124) for which the e*ltA* and *sat* genes were also detected in the metagenome, grouped closely with ETEC samples suggesting that this individual could have suffered from a co-infection with ETEC. Indeed, the estimated sequencing depth for the gene *eltA* (marker gene of ETEC) in this sample was 6X whereas the estimated sequencing depth for the *afaB-I* (marker gene for DAEC) was 2X. Therefore, we concluded that this sample was a co-infection with ETEC and its clustering position within ETEC samples was attributable to the high abundance of ETEC genotype. On the other hand, the metabolic profiling did not revealed differences among groups. Pathways related to *E. coli* virulence or antibiotic resistance as enterobactin biosynthesis (*p=4.07e-6*) and polymyxin resistance (*p=4.48e-5*) were observed to be significantly abundant among disease samples relative to controls, due to the significant increased signal of *E. coli* infection. However, after removing *E. coli* reads from the metagenomic libraries from DAEC and ETEC pathogroups and re-evaluating the metabolic profiles, no significant differences were detected. In addition, we estimated kmer-based dissimilarities among the *E. coli* and human-subtracted metagenome based on Mash distances (63). Our results in community dissimilarity based on k-mers of 25 nucleotides length or metabolic pathway abundance showed no clear separation among DAEC, ETEC and control groups. Therefore, we concluded that taxonomic, but not metabolic signatures occurred that discriminate the microbiomes associated with DAEC and ETEC infections.

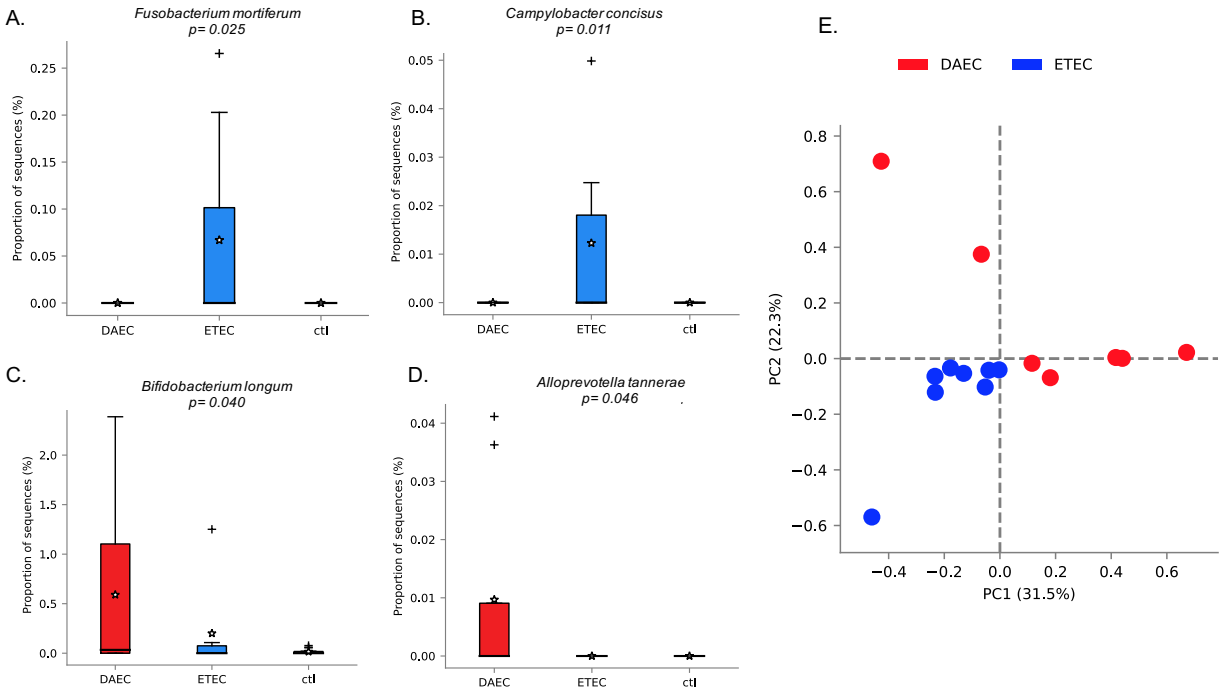


Figure 5.8. Differentially abundant taxa distinguishing between DAEC and ETEC infections. Differentially abundant species were reported if they had a corrected *p-value* ≤0.05 and an effect size (the magnitude of the difference between groups) ≥0.8 . Panels (A) and (B) shows the proportion of sequences assigned to *Fusobacterium mortiferum* and *Campylobacter concisus*, respectively. Similarly, panels (C) and (D) show the proportion of sequences assigned to *Bifidobacterium longum* and *Alloprevotella tannerae*, respectively. Panel (E) shows a principal component plot based on the taxonomic composition of each metagenome (annotated at the species level using clade-specific marker genes with MetaPhlAn2) after removing human and *E. coli* reads from the libraries

5.5 Conclusions and recommendations

In this study, we developed a bioinformatic approach to examine specific microbiome signatures in order to identify samples where *E. coli* was presumably the etiological agent causing diarrhea and assess pathotype-specific signatures. Our major findings revealed the presence of at least seven species differentially abundant between diarrhea and control samples, in general, with *Escherichia coli* and *Prevotella copri* being

the most discriminative. In addition, our results evidenced at least four species that discriminate DAEC from ETEC infections: *Fusobacterium mortiferum* and *Campylobacter concisus* were more abundant in ETEC infections while *Bifidobacterium longum* and *Alloprevotella tannerae* were significantly more enriched in DAEC infections. At the functional level, not major differences were detected. Our metagenomic approach has important implications for clinical microbiology and public health. Diagnostic testing for diarrheic pathogens has relied for decades on culture-based techniques that do not provide a quantitative picture of the pathogen or the response of the gut microbiome to the infection. Consequently, a significant fraction of diarrhea episodes remains un-diagnosed due to the challenges associated with the accurate detection of the pathogen responsible for the disease using traditional techniques. For example, the GEMS study reported that approximately 60% of the cases of moderate-to-severe diarrhea (MSD) in seven impoverished countries of Sub-Saharan Africa and South Asia remained undiagnosed because no known pathogen could be implicated using conventional diagnostic methods (65). Here, we show that metagenomic approaches can provide a higher resolution approach for the accurate detection of the pathogen responsible for the disease and for a better understanding of the disturbance of the gut microbiome by the infection. Specifically, the metagenomic picture consistently revealed that those diarrhea samples where *E. coli* was most likely the causative agent exhibited features consistent with a higher abundance of the pathogenic isolate compared to the control or the commensal population, high prevalence of virulence factors and/or toxins, including the metagenomic detection of the pathotype-specific marker gene for DAEC, ETEC and EPEC, and a reduced *E. coli* intra-population diversity compared with control samples. In comparison, about 50% of the diarrheal cases, which were identified by the traditional methodology to be (probably) caused by *E. coli* pathotypes, did not have strong metagenomic evidence for *E. coli* infection. The analysis for the other half of samples was not conclusive, which suggested that maybe other viral, bacterial or eukaryotic pathogens could have contributed to the development of disease. Alternatively, sample heterogeneity, e.g., our sample did not represent well the gut microbiome and thus, under-sampled the (potential) *E. coli*, could account for some of the samples with no strong metagenomic evidence of *E. coli* infection. Multiple replicate samples and high sampling volumes (e.g., 2-5 gr or more) should be used to avoid such sample heterogeneity issues. Regardless of the (probable) effects of sample heterogeneity, our finding collectively highlight that metagenomics is a promising diagnosis tool for

infectious diseases and the bioinformatic framework developed here can be applied to other pathogens of interest. Although several samples were excluded from the pathogen-specific signatures of the infection based on relatively low coverage observed in comparison with other positive samples or with controls, it is important to highlight that finding a 0.1-0.2X coverage still translates to a relative large number of cells *in situ* that potentially could produce successful infections. The average *E. coli* genome size is 5Mb and our metagenome libraries were in average ~2Gbs in size. Finding 0.2X coverage for an *E. coli* genome would represent ~0.05% of the total metagenome. Our extracted DNA came from an average of 10^{12} cells in total. Therefore, 0.05% of 10^{12} would be $5*10^{10}$ in total, which is still a large number of cells that could potentially cause a disease. Therefore, although metagenomics can advance clinical diagnosis, it has also limitations as for example, detection of low abundant microorganisms.

5.6 Acknowledgements

We thank to all participants from Ecuador (Quito, Esmeraldas, Borbón and the rural villages) who agreed to participate in the EcoZUR study and donated their specimens for the successful completion of this study. We would like to thank the local authorities of Ministry of Public Health for providing us access to the facilities in their hospitals. We would also like to thank Denys Tenorio, Mauricio Ayoví, Xavier Sanchez, Edison Puebla, and Kate Bohnert for their assistance with carrying out the field portions of the study. Funding for this study was provided by National Institute for Allergy and Infectious Diseases grant 1K01AI103544 and Colciencias through a doctoral fellowship to APG. The National Institute of Allergy and Infectious Diseases of the National Institutes of Health supported this work under Award Number K01AI103544. The content of this manuscript is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

5.7 References

1. Kotloff, K.L., 2017. The burden and etiology of diarrheal illness in developing countries. *Pediatric Clinics*, *64*(4), pp.799-814.

2. Lanata, C.F., Fischer-Walker, C.L., Olascoaga, A.C., Torres, C.X., Aryee, M.J. and Black, R.E., 2013. Global causes of diarrheal disease mortality in children< 5 years of age: a systematic review. *PloS one*, *8*(9), p.e72788.
3. Conway, T. and Cohen, P.S., 2015. Commensal and pathogenic Escherichia coli metabolism in the gut. *Microbiology spectrum*, *3*(3).
4. Kotloff, K.L., Nataro, J.P., Blackwelder, W.C., Nasrin, D., Farag, T.H., Panchalingam, S., Wu, Y., Sow, S.O., Sur, D., Breiman, R.F. and Faruque, A.S., 2013. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *The Lancet*, *382*(9888), pp.209-222.
5. Rossi, E., Cimdins, A., Lüthje, P., Brauner, A., Sjöling, Å., Landini, P. and Römling, U., 2018. "It'sa gut feeling"–Escherichia coli biofilm formation in the gastrointestinal tract environment. *Critical reviews in microbiology*, *44*(1), pp.1-30.
6. Kaper, J.B., Nataro, J.P. and Mobley, H.L., 2004. Pathogenic escherichia coli. *Nature reviews microbiology*, *2*(2), p.123.
7. Fleckenstein, J.M., 2013. Enterotoxigenic Escherichia coli. In *Escherichia coli (Second Edition)* (pp. 183-213).
8. Pasqua, M., Michelacci, V., Di Martino, M.L., Tozzoli, R., Grossi, M., Colonna, B., Morabito, S. and Prosseda, G., 2017. The intriguing evolutionary journey of enteroinvasive E. coli (EIEC) toward pathogenicity. *Frontiers in microbiology*, *8*, p.2390.
9. Kaur, P., Chakraborti, A. and Asea, A., 2010. Enteroaggregative Escherichia coli: an emerging enteric food borne pathogen. *Interdisciplinary perspectives on infectious diseases*, *2010*
10. Pearson, J.S., Giogha, C., Wong Fok Lung, T. and Hartland, E.L., 2016. The genetics of enteropathogenic Escherichia coli virulence. *Annual review of genetics*, *50*, pp.493-513.
11. Servin, A.L., 2014. Pathogenesis of human diffusely adhering Escherichia coli expressing Afa/Dr adhesins (Afa/Dr DAEC): current insights and future challenges. *Clinical microbiology reviews*, *27*(4), pp.823-869.
12. Smith, J.L., Fratamico, P.M. and Gunther IV, N.W., 2014. Shiga toxin-producing Escherichia coli. In *Advances in applied microbiology* (Vol. 86, pp. 145-197). Academic Press
13. Ridaura, V.K., Faith, J.J., Rey, F.E., Cheng, J., Duncan, A.E., Kau, A.L., Griffin, N.W., Lombard, V., Henrissat, B., Bain, J.R. and Muehlbauer, M.J., 2013. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science*, *341*(6150), p.1241214.
14. Everard, A., Belzer, C., Geurts, L., Ouwerkerk, J.P., Druart, C., Bindels, L.B., Guiot, Y., Derrien, M., Muccioli, G.G., Delzenne, N.M. and De Vos, W.M., 2013. Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proceedings of the National Academy of Sciences*, *110*(22), pp.9066-9071.
15. Kane, A.V., Dinh, D.M. and Ward, H.D., 2014. Childhood malnutrition and the intestinal microbiome. *Pediatric research*, *77*(1-2), p.256.
16. Morgan, X.C., Tickle, T.L., Sokol, H., Gevers, D., Devaney, K.L., Ward, D.V., Reyes, J.A., Shah, S.A., LeLeiko, N., Snapper, S.B. and Bousvaros, A., 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology*, *13*(9), p.R79.
17. Greenblum, S., Turnbaugh, P.J. and Borenstein, E., 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts

associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences*, *109*(2), pp.594-599.
18. Maloy, K.J. and Powrie, F., 2011. Intestinal homeostasis and its breakdown in inflammatory bowel disease. *Nature*, *474*(7351), p.298.
19. Pop, M., Walker, A.W., Paulson, J., Lindsay, B., Antonio, M., Hossain, M.A., Oundo, J., Tamboura, B., Mai, V., Astrovskaya, I. and Bravo, H.C., 2014. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome biology*, *15*(6), p.R76.
20. Monira, S., Nakamura, S., Gotoh, K., Izutsu, K., Watanabe, H., Alam, N.H., Nakaya, T., Horii, T., Ali, S.I., Iida, T. and Alam, M., 2013. Metagenomic profile of gut microbiota in children during cholera and recovery. *Gut pathogens*, *5*(1), p.1.
21. Forbes, J.D., Knox, N.C., Peterson, C.L. and Reimer, A.R., 2018. Highlighting clinical metagenomics for enhanced diagnostic decision-making: A step towards wider implementation. *Computational and Structural Biotechnology Journal*.
22. Toma, C., Lu, Y., Higa, N., Nakasone, N., Chinen, I., Baschkier, A., Rivas, M. and Iwanaga, M., 2003. Multiplex PCR assay for identification of human diarrheagenic Escherichia coli. *Journal of clinical microbiology*, *41*(6), pp.2669-2671.
23. Tornieporth, N.G., John, J., Salgado, K., de Jesus, P.A.U.L.O., Latham, E., Melo, M.C., Gunzburg, S.T. and Riley, L.W., 1995. Differentiation of pathogenic Escherichia coli strains in Brazilian children by PCR. *Journal of clinical microbiology*, *33*(5), pp.1371-1374.
24. Paton, A.W. and Paton, J.C., 1998. Detection and Characterization of Shiga ToxigenicEscherichia coli by Using Multiplex PCR Assays forstx 1, stx 2, eaeA, Enterohemorrhagic E. coli hlyA, rfb O111, andrfb O157. *Journal of Clinical Microbiology*, *36*(2), pp.598-602.
25. Le Bouguenec, C., Archambaud, M. and Labigne, A., 1992. Rapid and specific detection of the pap, afa, and sfa adhesin-encoding operons in uropathogenic Escherichia coli strains by polymerase chain reaction. *Journal of clinical microbiology*, *30*(5), pp.1189-1193.
26. Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K. and Schloss, P.D., 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and environmental microbiology*, pp.AEM-01043.
27. Cox, M.P., Peterson, D.A. and Biggs, P.J., 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC bioinformatics*, *11*(1), p.485.
28. Rodriguez-R, L.M., Gunturu, S., Harvey, W.T., Rosselló-Mora, R., Tiedje, J.M., Cole, J.R. and Konstantinidis, K.T., 2018. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. *Nucleic acids research*.
29. Peng, Y., Leung, H.C., Yiu, S.M. and Chin, F.Y., 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), pp.1420-1428.
30. Zhu, W., Lomsadze, A. and Borodovsky, M., 2010. Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, *38*(12), pp.e132-e132.
31. Luo, C., Rodriguez-r, L.M. and Konstantinidis, K.T., 2014. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic acids research*, *42*(8), pp.e73-e73.

32. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. and Holmes, S.P., 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods*, *13*(7), p.581.
33. Wang, Q., Garrity, G.M., Tiedje, J.M. and Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, *73*(16), pp.5261-5267.
34. Navas-Molina, J.A., Peralta-Sánchez, J.M., González, A., McMurdie, P.J., Vázquez-Baeza, Y., Xu, Z., Ursell, L.K., Lauber, C., Zhou, H., Song, S.J. and Huntley, J., 2013. Advancing our understanding of the human microbiome using QIIME. In *Methods in enzymology* (Vol. 531, pp. 371-444). Academic Press.
35. Faith, D.P. and Baker, A.M., 2006. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evolutionary bioinformatics*, *2*, p.117693430600200007.
36. Clarke, K.R., 1993. Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, *18*(1), pp.117-143.
37. Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H. and Oksanen, M.J., 2013. Package 'vegan'. *Community ecology package, version*, *2*(9).
38. Parks, D.H., Tyson, G.W., Hugenholtz, P. and Beiko, R.G., 2014. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics*, *30*(21), pp.3123-3124.
39. Wu, Y.W., Simmons, B.A. and Singer, S.W., 2015. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), pp.605-607.
40. Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), p.357.
41. Ramirez-Gonzalez, R.H., Bonnal, R., Caccamo, M. and MacLean, D., 2012. Bio-samtools: Ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments. *Source code for biology and medicine*, *7*(1), p.6.
42. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. and Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, pp.gr-186072.
43. Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P. and Tyson, G.W., 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature microbiology*, *2*(11), p.1533.
44. Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, *11*(1), p.119.
45. Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C. and Segata, N., 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods*, *12*(10), p.902.
46. Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J. and Cantarel, B.L., 2012. HUMAnN: The HMP unified metabolic analysis network. *PLoS Computational Biology*, *8*(6), p.e1002358.
47. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R. and Wu, C.H., 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, *23*(10), pp.1282-1288.
48. Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A. and Krummenacker, M., 2013.

The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*, *42*(D1), pp.D459-D471.
49. Ye, Y. and Doak, T.G., 2009. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS computational biology*, *5*(8), p.e1000465.
50. Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y. and Jin, Q., 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic acids research*, *33*(suppl_1), pp.D325-D328.
51. Pena-Gonzalez, A., Rodriguez-R, L.M., Marston, C.K., Gee, J.E., Gulvik, C.A., Kolton, C.B., Saile, E., Frace, M., Hoffmaster, A.R. and Konstantinidis, K.T., 2018. Genomic Characterization and Copy Number Variation of Bacillus anthracis Plasmids pXO1 and pXO2 in a Historical Collection of 412 Strains. *mSystems*, *3*(4), pp.e00065-18.
52. Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), pp.1792-1797.
53. Talavera, G. and Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, *56*(4), pp.564-577.
54. Price, M.N., Dehal, P.S. and Arkin, A.P., 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, *26*(7), pp.1641-1650.
55. Letunic, I. and Bork, P., 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research*, *44*(W1), pp.W242-W245.
56. Møller, A.K., Leatham, M.P., Conway, T., Nuijten, P.J., de Haan, L.A., Krogfelt, K.A. and Cohen, P.S., 2003. An Escherichia coli MG1655 lipopolysaccharide deep-rough core mutant grows and survives in mouse cecal mucus but fails to colonize the mouse large intestine. *Infection and immunity*, *71*(4), pp.2142-2152.
57. Leatham-Jensen, M.P., Frimodt-Møller, J., Adediran, J., Mokszycki, M.E., Banner, M.E., Caughron, J.E., Krogfelt, K.A., Conway, T. and Cohen, P.S., 2012. The Streptomycin-Treated Mouse Intestine Selects Escherichia coli envZ Missense Mutants that Interact with a Dense and Diverse Intestinal Microbiota. *Infection and immunity*, pp.IAI-06193.
58. Hoskins, L.C., 2018. Mucin degradation by enteric bacteria: ecological aspects and implications for bacterial attachment to gut mucosa. In *Attachment of organisms to the gut mucosa*(pp. 51-68). CRC Press.
59. Rodriguez-r, L.M. and Konstantinidis, K.T., 2013. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics*, *30*(5), pp.629-635.
60. Huang, A.D., Luo, C., Pena-Gonzalez, A., Weigand, M.R., Tarr, C.L. and Konstantinidis, K.T., 2017. Metagenomics of two severe foodborne outbreaks provides diagnostic signatures and signs of coinfection not attainable by traditional methods. *Applied and environmental microbiology*, *83*(3), pp.e02577-16.
61. Apperloo-Renkema, H.Z., Van der Waaij, B.D. and Van der Waaij, D., 1990. Determination of colonization resistance of the digestive tract by biotyping of Enterobacteriaceae. *Epidemiology & Infection*, *105*(2), pp.355-361.
62. Qiu, F.Z., Shen, X.X., Li, G.X., Zhao, L., Chen, C., Duan, S.X., Guo, J.Y., Zhao, M.C., Yan, T.F., Qi, J.J. and Wang, L., 2018. Adenovirus associated with acute diarrhea: a case-control study. *BMC infectious diseases*, *18*(1), p.450.

63. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S. and Phillippy, A.M., 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology*, *17*(1), p.132.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

A.1 Draft genome sequence of *Bacillus cereus* LA2007, a human-pathogenic isolate harboring anthrax-like plasmids.

Partially reproduced with permission from Angela Pena-Gonzalez, Chung K. Marston, Luis M. Rodriguez-R, Cari B. Kolton, Julia Garcia-Diaz, Amanda Theppote, Michael Frace, Konstantinos T. Konstantinidis and Alex R. Hoffmaster. Genome Announc. 2017, 5(16), pil:e00181-17.

A.1.1 Summary

We present the genome sequence of *Bacillus cereus* LA-2007, a strain isolated in 2007 from a fatal pneumonia case in Louisiana. Sequence-based genome analysis revealed that LA-2007 carries a plasmid highly similar to *Bacillus anthracis* pXO1, including the genes responsible for production and regulation of anthrax toxin.

A.1.2 Genome announcement

Bacillus cereus strains encoding genetic determinants that confer pathogenic capabilities similar to those found on *B. anthracis* have been recently described (1-5). In *B. anthracis*, virulence determinants are encoded on two large plasmids, pXO1 and pXO2, including the genes necessary to produce the anthrax toxin (*lef*, *cya* and *pagA*) and a D-polyglutamic acid capsule (*capBCADE*) allowing the pathogen to evade host immune response. Plasmids similar to pXO1 and pXO2 have been found in a number of previously sequenced *B. cereus sensu lato* strains causing anthrax-like diseases (1-5). It has not been determined yet whether or not lateral transfer has occurred between members of *B. cereus* and *B. anthracis*, the direction of such a transfer, or the phylogenetic relationships of the strains based on the plasmid sequences. To gain deeper insights into the evolution of *B. cereus* strains causing anthrax-like diseases, we

present the draft genome of LA-2007, a bacterial pathogen isolated from a fatal pneumonia case in a female welder from Louisiana.

DNA was extracted from overnight culture on trypticase soy agar with 5% sheep blood using the Maxwell 16 Promega Instrument and sequencing was performed on an Illumina GAIIx platform. Read quality control, assembly and annotation were performed as described in Tsementzi (2016) (6). The assembled genome of LA-2007 consisted of 67 contigs with a GC content of 35% and a total genome size of 5,224,740 bp. The estimated percent of completeness and contamination were 99.15% and 0.28%, respectively. The genome was predicted to contain a total of 5,777 putative protein-coding sequences, 15 rRNA operons and 57 tRNAs genes. The calculated Average Nucleotide Identity (ANI) (**7**) of LA-2007 against *B. anthracis* Ames was 94.76%, and against a set of *B. cereus* strains previously reported to be associated with severe pneumonia similar to LA-2007, i.e., strains G9241 (2), 03BB87 (**3**) and BcFL2013 (1), was 99.99% in all three cases, indicating that all four *B. cereus* strains are part of the same clonal complex.

Characterization of plasmid gene content showed that the genome of *B. cereus* LA-2007 contains a pXO1-like plasmid assembled in eight contigs, each one of them showing at least 99.70% ANI and 80% gene coverage with *B. anthracis* Ames pXO1. The latter results suggest that LA-2007 (or its ancestors) might have acquired the pXO1 plasmid horizontally from *B. anthracis* (since ANI of the plasmid is higher than that of the main chromosome), although detailed phylogenetic analysis will be required to robustly test this finding. Comparison with plasmid pBCXO1 from strains G9241, 03BB87 and BcFL2013 revealed an ANI of ≥99.99% in all three cases, similar to the chromosomes mentioned above. A plasmid homologous to pBc210 reported in *B. cereus* G9241, was also identified in LA-2007. The latter plasmid was assembled in 9 contigs and showed 99.98% ANI when compared to the plasmid of G9241. Anthrax toxin genes as well as complete genes for production of hyaluronic acid synthases (*hasACB*) and exo-polysaccharides (*bpsH-X)* were also identified in LA-2007, suggesting its ability to produce protective capsules.

A.1.3 Nucleotide sequence accession numbers.

This Whole Genome Shotgun project has been deposited at GenBank under the accession MUBB00000000. The version described in this paper is version MUBB01000000.

A.1.4 Acknowledgements

This work was supported by a doctoral scholarship to A.P-G from Colciencias, Colombian Administrative Department for Science, Technology and Innovation and by the C.D.C. award # RF023.

A.1.5 References

1. **Marston CK, Ibrahim H, Lee P, Churchwell G, Gumke M, Stanek D, Gee JE, Boyer AE, Gallegos-Candela M, Barr JR**. 2016. Anthrax toxin-expressing *Bacillus cereus* isolated from an anthrax-like eschar. PloS one 11:e0156987

2. **Hoffmaster AR, Ravel J, Rasko DA, Chapman GD, Chute MD, Marston CK, De BK, Sacchi CT, Fitzgerald C, Mayer LW**. 2004. Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. Proceedings of the National Academy of Sciences of the United States of America 101:8449-8454.

3. **Hoffmaster AR, Hill KK, Gee JE, Marston CK, De BK, Popovic T, Sue D, Wilkins PP, Avashia SB, Drumgoole R**. 2006. Characterization of *Bacillus cereus* isolates associated with fatal pneumonias: strains are closely related to Bacillus anthracis and harbor B. anthracis virulence genes. Journal of clinical microbiology 44:3352-3360.

4. **Antonation KS, Grützmacher K, Dupke S, Mabon P, Zimmermann F, Lankester F, Peller T, Feistner A, Todd A, Herbinger I**. 2016. Bacillus cereus Biovar Anthracis Causing Anthrax in Sub-Saharan Africa—Chromosomal Monophyly and Broad Geographic Distribution. PLoS Negl Trop Dis 10:e0004923.

5. **Klee SR, Özel M, Appel B, Boesch C, Ellerbrok H, Jacob D, Holland G, Leendertz FH, Pauli G, Grunow R**. 2006. Characterization of Bacillus anthracis-like bacteria isolated from wild great apes from Cote d'Ivoire and Cameroon. Journal of bacteriology 188:5333-5344.

6. **Tsementzi D, Wu J, Deutsch S, Nath S, Rodriguez-r LM, Burns AS, Ranjan P, Sarode N, Malmstrom RR, Padilla CC**. 2016. SAR11 bacteria linked to ocean anoxia and nitrogen loss. Nature 536:179-183.

7. **Rodriguez-R LM, Konstantinidis KT**. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints.

APPENDIX B

SUPPLEMENTARY TABLES FOR CHAPTER 4

Table B.1. Genes and primers used in the PCR identification of *E. coli* pathotypes

E. coli	Gene	Primer sequence 5 – 3'	Size (bp)
EAEC	*aggR*[a]	5'GTATACACAAAAGAAGGAAGC3'	254
		5'ACAGAATCGTCAGCATCAGC3'	
ETEC	*eltA*[b]	5'GCGACAAATTATACCGTGCT3'	708
		5'CCGAATTCTGTTATATATGT3'	
	sta[c]	5'CTGTATTGTCTTTTTCACCT3'	182
		5'GCACCCGGTACAAGCAGGAT3'	
EPEC	*bfp*[d]	5'CAATGGTGCTTGCGCTTGCT3'	324
		5'GCCGCTTTATCCAACCTGGT3'	
	eaeA[e]	5'GACCCGGCACAAGCATAAGC3'	384
		5'CCACCTGCAGCAACAAGAGG3'	
EIEC	*ipaH*[f]	5'GCTGGAAAAACTCAGTGCCT3'	424
		5'CCAGTCCGTAAATTCATTCT3'	
DAEC	*afa*[g]	5'GCTGGGCAGCAAACTGATAACTCTC3'	750
		5'CATCAAGCTCTTTGTTCGTCCGCCG3'	
STEC	*stx1*[h]	5'ATAAATCGCCATTCGTTGACTAC3'	180
		5'AGAACGCCCACTGAGATCATC3'	
	stx2[h]	5'TCGCCAGTTATCTGACATTCTG3'	255

Table B.2. Clinical and demographic information of individuals from Ecuador enrolled in the EcoZUR study where *E. coli* isolate was cultured.

Sample ID	*Location*	*Clinical*	*Pathotype-PCR*	*Pathotype-BIOINF*	*Gender*	*Age [months]*
B001 5	rural	diarrhea	DAEC	DAEC	male	14
B100 5	rural	asymptom	DAEC	DAEC	female	325
B101 1	rural	asymptom	DAEC	DAEC	male	6
B108 3	rural	diarrhea	EAEC	EAEC	male	11
B109 1	rural	diarrhea	ETEC	ETEC	female	70
B118 2	rural	asymptom	DAEC	DAEC	female	657
B119 1	rural	diarrhea	EAEC	EAEC	male	11
B119 2	rural	diarrhea	EPECa	EAEC	male	11
B12 1	rural	diarrhea	DAEC	DAEC	male	22
B124 2	rural	asymptom	EPECa	NA	female	50
B135 1	rural	diarrhea	EIEC	NA	female	52
B141 2	rural	asymptom	EAEC	EAEC	female	55
B143 4	rural	asymptom	EIEC	EIEC	male	11
B145 4	rural	asymptom	EAEC	EAEC	male	9
B147 1	rural	asymptom	DAEC	DAEC	male	15
B169 1	rural	asymptom	EAEC	EAEC	female	9
B188 1	rural	asymptom	EAEC	EAEC	female	124
B196 1	rural	diarrhea	EPECa	EPECa	female	394
B200 2	rural	diarrhea	ETEC	ETEC	female	31
B201 1	rural	diarrhea	EPECa	EIEC	male	52
B201 3	rural	diarrhea	EAEC	EAEC	male	52
B201 5	rural	diarrhea	EAEC	EAEC	male	52
B202 2	rural	diarrhea	EAEC	EAEC	male	184
B207 2	rural	diarrhea	DAEC	DAEC	female	147
B215 5	rural	Diarrhea	EAEC	EAEC	female	20

Table B.2 continued

B225 3	rural	asymptom	EPECa	EPECa	female	36
B226 1	rural	diarrhea	DAEC	DAEC	male	773
B228 2	rural	diarrhea	EPECa	NA	female	68
B231 1	rural	asymptom	EPECt	EPECt	female	11
B231 2	rural	asymptom	EPECa	EPECa	female	11
B234 1	rural	asymptom	DAEC	DAEC	male	353
B235 3	rural	diarrhea	EPECt	EPECt	female	1024
B24 1	rural	diarrhea	DAEC	DAEC	male	27
B244 3	rural	diarrhea	ETEC	ETEC	male	15
B246 1	rural	diarrhea	EPECa	EPECa	male	368
B252 1	rural	diarrhea	EPECa	NA	male	418
B255 1	rural	diarrhea	ETEC	ETEC	female	21
B259 1	rural	asymptom	EPECa	EPECa	female	79
B262 1	rural	diarrhea	EIEC	EIEC	female	795
B270 2	rural	diarrhea	ETEC	ETEC	male	531
B274 2	rural	diarrhea	DAEC	DAEC	female	16
B280 4	rural	asymptom	EAEC	EAEC	male	479
B28 1	rural	diarrhea	EAEC	NA	missing	missing
B295 2	rural	diarrhea	ETEC	ETEC	male	27
B309 1	rural	diarrhea	EIEC	EIEC	female	42
B311 2	rural	diarrhea	EPECa	NA	female	800
B312 4	rural	asymptom	EPECa	EPECa	female	18
B313 5	rural	asymptom	EPECa	NA	female	399
B323 2	rural	asymptom	EPECa	NA	female	14
B327 1	rural	asymptom	EPECa	NA	male	145
B329 2	rural	asymptom	DAEC	DAEC	male	9
B36 1	rural	asymptom	EPECa	EAEC	female	19
B37 1	rural	diarrhea	DAEC	NA	female	530

Table B.2 continued

B37 2	rural	diarrhea	DAEC	NA	female	530
B37 3	rural	diarrhea	DAEC	NA	female	530
B37 6	rural	diarrhea	DAEC	NA	female	530
B42 1	rural	diarrhea	DAEC	NA	male	365
B42 3	rural	diarrhea	DAEC	NA	male	365
B45 2	rural	diarrhea	ETEC	ETEC	female	25
B46 1	rural	asymptom	ETEC	ETEC	male	17
B62 5	rural	diarrhea	ETEC	ETEC	female	18
B66 1	rural	diarrhea	EAEC	EAEC	male	8
B66 4	rural	diarrhea	ETEC	EAEC	male	8
B68 1	rural	diarrhea	ETEC	ETEC	male	27
B69 1	rural	diarrhea	EPECa	NA	female	24
B75 4	rural	diarrhea	EPECa	NA	male	14
B75 5	rural	diarrhea	DAEC	DAEC	male	14
B84 2	rural	diarrhea	ETEC	ETEC	female	13
B84 3	rural	diarrhea	EIEC	EIEC	female	13
B88 3	rural	diarrhea	EIEC	EIEC	female	15
B89 1	rural	diarrhea	DAEC	DAEC	female	13
B95 1	rural	diarrhea	DAEC	DAEC	female	6
B99 5	rural	asymptom	EPECa	EPECa	male	95
C101 1	rural	diarrhea	ETEC	ETEC	male	50
C102 1	rural	asymptom	EPECa	EPECa	male	30
C110 5	rural	asymptom	EPECa	EPECa	female	47
C1 2	rural	diarrhea	ETEC	ETEC	missing	missing
C13 2	rural	diarrhea	EPECa	NA	male	30
C14 2A	rural	diarrhea	DAEC	DAEC	female	113
C14 2B	rural	diarrhea	DAEC	DAEC	female	113
C21_2	rural	asymptom	EAEC	EAEC	female	35

Table B.2 continued

C21 4	rural	asymptom	EAEC	EAEC	female	35
C23 1	rural	asymptom	EAEC	EAEC	male	15
C25 2	rural	asymptom	EIEC	EIEC	female	183
C26 2	rural	asymptom	EPECa	EPECa	male	40
C32 3	rural	diarrhea	ETEC	NA	male	14
C33 3	rural	diarrhea	EAEC	EAEC	male	52
C34 2	rural	diarrhea	DAEC	DAEC	male	23
C36 1	rural	diarrhea	DAEC	DAEC	male	15
C38 1	rural	diarrhea	ETEC	NA	male	25
C46 4	rural	asymptom	ETEC	ETEC	male	31
C50 4	rural	asymptom	DAEC	DAEC	female	446
C70 1	rural	asymptom	EPECa	EPECa	male	16
C71 2	rural	asymptom	DAEC	DAEC	male	17
C72 3	rural	diarrhea	EAEC	EAEC	missing	missing
C79 1	rural	diarrhea	DAEC	DAEC	male	14
C80 4	rural	asymptom	EAEC	EAEC	male	5
C85 1	rural	diarrhea	ETEC	NA	female	40
C86 4	rural	asymptom	EAEC	EPECa	female	15
C87 3	rural	diarrhea	ETEC	NA	female	325
C9 2	rural	diarrhea	EPECa	EPECa	male	91
E100 4	urban	diarrhea	DAEC	DAEC	female	50
E100 5	urban	diarrhea	DAEC	DAEC	female	50
E101 1	urban	asymptom	DAEC	DAEC	male	37
E106 1	urban	diarrhea	EAEC	EAEC	male	19
E119 5	urban	asymptom	DAEC	DAEC	female	40
E124 5	urban	diarrhea	EAEC	EAEC	female	14
E124 6	urban	diarrhea	DAEC	DAEC	female	14
E129_3	urban	diarrhea	EPECa	EPECt	male	6

Table B.2 continued

E130 1	urban	asymptom	ETEC	EAEC	male	20
E132 5	urban	asymptom	DAEC	EAEC	male	15
E135 2	urban	diarrhea	EAEC	EAEC	male	124
E135 5	urban	diarrhea	EAEC	EAEC	male	124
E135 6	urban	diarrhea	DAEC	DAEC	male	124
E139 2	urban	asymptom	ETEC	ETEC	female	19
E13	urban	diarrhea	DAEC	DAEC	male	166
E140	urban	diarrhea	DAEC	DAEC	male	355
E143 1	urban	asymptom	EPECa	NA	male	443
E144	urban	diarrhea	EPECa	EPECa	female	835
E158	urban	diarrhea	DAEC	DAEC	male	13
E162	urban	diarrhea	EPECa	EPECa	female	78
E16 6	urban	asymptom	DAEC	DAEC	male	66
E166	urban	diarrhea	DAEC	DAEC	male	125
E167 5	urban	asymptom	DAEC	DAEC	female	99
E170	urban	diarrhea	DAEC	DAEC	male	225
E173	urban	diarrhea	DAEC	DAEC	male	934
E175 4	urban	asymptom	EAEC	EAEC	female	28
E175 5	urban	asymptom	ETEC	ETEC	female	28
E177 6	urban	diarrhea	DAEC	DAEC	male	9
E179 6	urban	asymptom	DAEC	DAEC	male	14
E184 3	urban	diarrhea	ETEC	NA	male	16
E187 2	urban	diarrhea	EPECt	EPECt	female	21
E188 2	urban	diarrhea	DAEC	DAEC	female	157
E192	urban	diarrhea	DAEC	DAEC	male	11
E195 3	urban	diarrhea	ETEC	NA	female	16
E195 5	urban	diarrhea	EPECa	EPECa	female	16
E197 5	urban	asymptom	DAEC	DAEC	male	16

Table B.2 continued

E20 1	urban	asymptom	EAEC	EAEC	male	24
E205	urban	diarrhea	EPECa	NA	female	21
E212 4	urban	diarrhea	DAEC	DAEC	female	796
E21 3	urban	asymptom	missing	NA	male	17
E215 4	urban	asymptom	ETEC	NA	male	28
E218	urban	diarrhea	DAEC	DAEC	female	190
E228 3	urban	asymptom	EAEC	EAEC	female	7
E230 4	urban	diarrhea	DAEC	DAEC	male	14
E236 4	urban	asymptom	EAEC	EAEC	female	14
E238 3	urban	asymptom	EAEC	EAEC	female	5
E238 5	urban	asymptom	DAEC	DAEC	female	5
E240 5	urban	asymptom	EPECa	EPECa	female	3
E26	urban	diarrhea	DAEC	DAEC	female	51
E27	urban	diarrhea	DAEC	DAEC	male	15
E28	urban	diarrhea	DAEC	DAEC	male	162
E29 1	urban	diarrhea	ETEC	ETEC	male	232
E33 4	urban	asymptom	EPECa	EPECa	female	161
E34	urban	diarrhea	DAEC	DAEC	male	292
E35 4	urban	asymptom	EPECa	EPECa	male	23
E43	urban	diarrhea	DAEC	DAEC	male	368
E44 6	urban	diarrhea	DAEC	DAEC	male	24
E55 5	urban	asymptom	ETEC	ETEC	male	16
E55 6	urban	asymptom	DAEC	DAEC	male	16
E5 5	urban	diarrhea	EPECa	EPECa	male	15
E57	urban	diarrhea	DAEC	DAEC	female	140
E58 3	urban	diarrhea	DAEC	DAEC	male	6
E61 3	urban	diarrhea	ETEC	NA	male	10
E64_6	urban	missing	missing	NA	missing	missing

Table B.2 continued

E67 5	urban	diarrhea	DAEC	DAEC	male	8
E70 3	urban	asymptom	EAEC	EAEC	female	105
E71 6	urban	diarrhea	EIEC	EIEC	male	336
E71	urban	diarrhea	EPECa	EPECa	male	336
E72 3	urban	diarrhea	EPECa	NA	male	97
E72	urban	diarrhea	DAEC	DAEC	male	97
E76 1	urban	asymptom	EAEC	EAEC	male	700
E84	urban	diarrhea	EPECa	NA	male	356
E88 3	urban	diarrhea	DAEC	DAEC	male	84
E88 4	urban	diarrhea	EAEC	EAEC	male	84
E89 4	urban	diarrhea	EAEC	EAEC	male	147
E92 5	urban	asymptom	DAEC	DAEC	male	9
E99 3	urban	asymptom	DAEC	DAEC	female	483
Q106	urban	asymptom	ETEC	ETEC	male	92
Q108	urban	diarrhea	EIEC	EIEC	female	126
Q128	urban	diarrhea	EIEC	EIEC	female	238
Q132	urban	asymptom	ETEC	ETEC	male	43
Q142	urban	diarrhea	DAEC	DAEC	female	402
Q145	urban	asymptom	DAEC	DAEC	female	365
Q147	urban	diarrhea	DAEC	DAEC	female	379
Q16 1	urban	diarrhea	ETEC	ETEC	female	720
Q186	urban	asymptom	EAEC	EAEC	male	106
Q196	urban	diarrhea	DAEC	DAEC	male	59
Q199	urban	asymptom	DAEC	DAEC	female	10
Q21 2	urban	diarrhea	DAEC	DAEC	female	570
Q212	urban	asymptom	EAEC	EAEC	male	152
Q223	urban	diarrhea	DAEC	DAEC	female	483
Q23 1	urban	diarrhea	DAEC	DAEC	female	498

Table B.2 continued

Q233	urban	diarrhea	EPECa	EPECa	male	32
Q240	urban	diarrhea	DAEC	DAEC	male	530
Q243	urban	diarrhea	DAEC	DAEC	female	518
Q245 2	urban	asymptom	DAEC	DAEC	male	44
Q249	urban	asymptom	EPECa	EPECa	male	56
Q250 1	urban	asymptom	EAEC	EAEC	female	0
Q253	urban	asymptom	EPECa	EPECa	male	32
Q270	urban	diarrhea	EPECa	EPECa	female	242
Q27 1	urban	asymptom	EPECa	EPECa	male	27
Q275	urban	asymptom	EPECa	EPECa	female	594
Q279 1	urban	diarrhea	EIEC	EIEC	male	146
Q282	urban	asymptom	EPECa	EPECa	female	30
Q284	urban	asymptom	EPECa	EPECa	female	16
Q288	urban	asymptom	ETEC	ETEC	male	99
Q289	urban	asymptom	EPECa	NA	female	360
Q294	urban	diarrhea	ETEC	ETEC	female	18
Q295	urban	diarrhea	ETEC	ETEC	female	947
Q300	urban	diarrhea	EPECa	NA	male	59
Q307 1	urban	asymptom	DAEC	DAEC	female	0
Q308	urban	diarrhea	DAEC	DAEC	female	23
Q310	urban	diarrhea	DAEC	DAEC	female	16
Q312 1	urban	asymptom	DAEC	DAEC	male	7
Q31 6	urban	diarrhea	EIEC	EIEC	female	325
Q33 6	urban	diarrhea	EIEC	EIEC	female	70
Q34 6	urban	diarrhea	EIEC	EIEC	female	239
Q35 1	urban	diarrhea	DAEC	DAEC	female	522
Q37 4	urban	asymptom	EPECa	EPECa	male	408
Q38 1	urban	diarrhea	ETEC	ETEC	male	467

Table B.2 continued

Q40	urban	diarrhea	DAEC	DAEC	male	465
Q49	urban	diarrhea	DAEC	NA	male	21
Q51	urban	diarrhea	DAEC	DAEC	male	48
Q53	urban	diarrhea	ETEC	ETEC	female	61
Q56	urban	diarrhea	DAEC	DAEC	female	17
Q65	urban	diarrhea	DAEC	DAEC	female	24
Q71 1	urban	diarrhea	ETEC	ETEC	male	13
Q85 1	urban	diarrhea	DAEC	DAEC	male	128
Q86	urban	asymptom	EPECa	EPECa	male	22
Q87	urban	asymptom	ETEC	ETEC	male	22
Q89	urban	diarrhea	EAEC	EAEC	male	380
Q91	urban	diarrhea	DAEC	DAEC	male	313
R104 2	rural	asymptom	EPECa	EPECa	male	111
R109 1	rural	asymptom	EPECa	EPECa	female	19
R110 5	rural	asymptom	EPECa	NA	male	8
R113 1	rural	diarrhea	DAEC	DAEC	male	6
R113 3	rural	diarrhea	ETEC	ETEC	male	6
R116 2	rural	asymptom	ETEC	ETEC	male	204
R119 3	rural	asymptom	ETEC	ETEC	male	94
R12 1	rural	asymptom	ETEC	NA	female	22
R122 4	rural	diarrhea	EAEC	EAEC	male	134
R127 3	rural	diarrhea	EPECa	EPECa	male	91
R137 1	rural	diarrhea	EPECa	NA	male	28
R138 1	rural	asymptom	EPECa	NA	female	176
R17 2	rural	diarrhea	ETEC	ETEC	male	453
R36 1	rural	diarrhea	EAEC	EAEC	male	21
R42 2	rural	diarrhea	DAEC	DAEC	male	362
R43_2	rural	diarrhea	EPECa	EPECa	male	32

Table B.2 continued

R46 3	rural	diarrhea	ETEC	ETEC	male	60
R55 1	rural	asymptom	EPECa	EPECa	female	191
R56 3	rural	asymptom	EAEC	EAEC	male	190
R58 1	rural	asymptom	EPECa	NA	male	160
R60 2	rural	asymptom	EPECa	NA	female	461
R66 4	rural	diarrhea	ETEC	ETEC	female	99
R66 5	rural	diarrhea	ETEC	ETEC	female	99
R67 3	rural	asymptom	DAEC	DAEC	male	130
R83 3	rural	asymptom	EAEC	EAEC	male	28
R84 2	rural	diarrhea	EPECa	NA	male	15
R8 4	rural	asymptom	DAEC	DAEC	male	21
R85 2	rural	asymptom	ETEC	ETEC	female	25
R86 1	rural	diarrhea	DAEC	DAEC	female	961
R92_5	rural	asymptom	EAEC	EAEC	male	41

Table B.3. Multilocus sequence analysis in the collection of *E. coli* isolates.

strain ID	***adk***	***fumC***	***gyrB***	***icd***	***mdh***	***purA***	***recA***	**ST**	**Phylogroup**
B001_5	6	11	4	8	8	78	2	1312	A
B100_5	53	40	47	13	36	28	29	131	B2
B101_1	35	37	29	25	4	5	73	405	D
B108_3	6	4	5	26	7	8	14	200	B1
B109_1	9	6	162	131	24	8	7	3857	B1
B118_2	10	11	4	8	8	8	2	10	A
B119_1	10	11	4	8	8	8	2	10	A

Table B.3 continued

B119_2	154	187	22	52	130	129	4	973	D
B12_1	53	40	47	13	36	28	29	131	B2
B124_2	6	11	4	10	7	8	6	93	A
B135_1	6	4	14	16	24	5	14	1049	B1
B141_2	6	6	5	136	9	7	7	678	B1
B143_4	6	4	7	1	11	3	56	270	B1
B145_4	10	11	4	8	8	8	2	10	A
B147_1	21	35	61	52	5	5	4	394	D
B169_1	35	132	2	27	37	5	4	501	D
B188_1	10	11	4	8	8	8	2	10	A
B196_1	16	4	12	16	9	7	7	21	B1
B200_2	9	6	33	131	24	8	7	641	B1
B201_1	6	4	7	1	11	3	56	270	B1
B201_3	10	11	4	8	8	8	2	10	A
B201_5	10	11	4	8	8	8	2	10	A
B202_2	9	23	64	18	11	8	6	278	B1
B207_2	10	11	4	8	8	9	2	4238	A
B215_5	10	11	4	1	8	8	2	34	A
B225_3	6	4	4	16	43	8	6	603	B1
B226_1	4	26	2	25	5	5	19	38	D
B228_2	21	35	61	52	5	5	4	394	D
B231_1	13	21	13	22	17	14	15	28	B2
B231_2	6	4	14	16	24	8	14	155	B1

Table B.3 continued

B234_1	10	11	4	1	8	9	2	227	A
B235_3	13	21	13	22	17	14	15	28	B2
B24_1	6	11	4	8	8	8	2	48	A
B244_3	6	5	4	8	8	8	2	4	A
B246_1	10	11	4	8	8	8	49	752	A
B252_1	6	11	4	8	8	8	2	48	A
B255_1	11	7	4	8	12	8	2	na	A
B259_1	109	65	5	1	9	13	14	517	B1
B262_1	6	4	7	1	11	3	56	270	B1
B270_2	6	11	4	8	331	78	2	3931	A
B274_2	10	11	4	1	8	9	2	227	A
B28_1	6	4	15	246	11	8	6	8097	B2
B280_4	10	7	4	8	12	8	2	746	A
B295_2	49	4	44	141	11	35	7	na	A
B309_1	11	63	7	1	14	7	7	152	*Shigella*
B311_2	38	39	9	13	17	44	34	7804	B2
B312_4	10	11	4	1	8	66	2	378	A
B313_5	6	19	124	16	11	8	105	na	B2
B323_2	6	1037	3	18	11	122	71	na	B2
B327_1	10	11	272	8	8	13	2	na	A
B329_2	10	11	4	8	8	8	2	10	A
B36_1	10	11	4	8	8	8	2	10	A
B37_1	10	11	4	8	8	140	2	1078	A

Table B.3 continued

B37_2	111	11	4	8	8	8	2	559	A
B37_3	6	23	15	225	11	8	7	na	B2
B37_6	82	749	491	595	458	425	406	5792	B2
B42_1	6	11	4	10	7	8	6	93	A
B42_3	6	11	4	10	7	8	6	93	A
B45_2	6	5	4	8	8	8	2	4	A
B46_1	6	5	4	8	8	8	2	4	A
B62_5	6	5	4	8	8	8	2	4	A
B66_1	10	11	4	8	8	8	2	10	A
B66_4	10	11	4	8	8	8	2	10	A
B68_1	9	6	162	131	24	8	7	3857	B1
B69_1	10	11	4	8	8	13	73	617	A
B75_4	53	40	47	13	36	28	29	131	B2
B75_5	4	26	2	25	5	5	19	38	D
B84_2	9	6	162	131	24	8	7	3857	B1
B84_3	6	4	7	1	11	3	56	270	B1
B88_3	8	7	1	1	10	8	6	6	A
B89_1	10	11	4	8	8	8	7	44	A
B95_1	4	26	2	25	5	5	19	38	D
B99_5	6	135	12	16	9	7	7	2836	B1
C1_2	10	27	5	10	96	8	2	4121	A
C101_1	6	11	551	713	331	1	156	na	B1
C102_1	10	11	4	1	8	66	2	378	A

Table B.3 continued

C110_5	76	24	244	89	17	14	79	4554	B2
C13_2	24	11	4	8	8	8	2	43	A
C14_2A	4	26	2	25	5	5	19	38	D
C14_2B	35	37	29	25	4	5	73	405	D
C21_2	10	11	4	8	8	8	2	10	A
C21_4	10	11	4	8	8	8	2	10	A
C23_1	10	11	4	1	8	8	2	34	A
C25_2	6	4	7	1	11	3	56	270	B1
C26_2	109	65	5	1	9	13	14	517	B1
C32_3	21	35	27	6	5	5	4	69	D
C33_3	6	4	12	1	9	2	7	295	B1
C34_2	6	4	566	1	20	13	7	na	B1
C36_1	6	4	12	1	20	13	7	23	B1
C38_1	21	35	27	6	5	5	4	69	D
C46_4	6	5	4	8	8	8	2	4	A
C50_4	83	627	245	404	91	17	181	5148	D
C70_1	142	67	13	150	91	115	11	788	B2
C71_2	35	37	29	25	4	5	73	405	D
C72_3	35	132	2	27	37	5	4	501	D
C79_1	10	7	4	8	12	8	2	746	A
C80_4	10	11	4	1	8	8	2	34	A
C85_1	41	6	5	16	11	8	7	94	B1
C86_4	6	68	5	1	24	8	7	na	B2

Table B.2 continued

C87_3	6	11	4	8	8	8	557	7684	A
C9_2	6	4	12	16	9	7	7	29	B1
E100_4	4	26	2	25	5	5	19	38	D
E100_5	83	627	245	404	91	17	181	5148	D
E101_1	6	11	4	8	8	78	2	1312	A
E106_1	6	4	5	26	7	8	14	200	B1
E119_5	10	11	4	8	8	8	2	10	A
E124_5	6	4	5	26	7	8	14	200	B1
E124_6	14	14	10	200	17	7	10	1193	B2
E129_3	6	7	5	1	8	18	2	206	A
E13	53	40	47	13	36	28	29	131	B2
E130_1	21	35	27	6	5	5	4	69	D
E132_5	10	11	4	8	8	8	2	10	A
E135_2	10	11	4	8	8	8	2	10	A
E135_5	10	11	4	8	8	8	2	10	A
E135_6	83	627	245	404	91	17	181	5148	D
E139_2	10	27	5	10	96	8	2	4121	A
E140	21	35	27	6	5	5	4	69	D
E143_1	10	11	183	8	8	8	2	1141	A
E144	6	4	5	26	20	8	14	40	B1
E158	4	26	2	25	5	5	19	38	D
E16_6	4	26	2	25	5	5	19	38	D
E162	129	311	13	150	91	12	25	6125	B2

Table B.3 continued

E166	35	37	29	25	4	5	73	405	D
E167_5	10	11	4	8	8	8	2	10	A
E170	53	40	47	13	36	28	29	131	B2
E173	21	35	61	52	5	5	4	394	D
E175_4	6	4	12	1	9	2	7	295	B1
E175_5	6	95	3	274	9	8	2	2602	B1
E177_6	53	40	47	13	36	28	29	131	B2
E179_6	53	na	47	13	36	28	29	na	B2
E184_3	274	4	96	1	24	8	6	2332	A
E187_2	13	21	13	22	17	14	15	28	B2
E188_2	83	627	245	404	91	17	181	5148	D
E192	4	26	2	25	5	5	19	38	D
E195_3	41	6	5	16	11	8	7	94	B1
E195_5	142	67	13	150	91	115	11	788	B2
E197_5	35	183	29	25	4	5	73	964	D
E20_1	10	11	4	8	8	8	2	10	A
E205	21	35	61	52	5	5	4	394	D
E21_3	6	11	4	10	7	8	6	93	A
E212_4	4	26	2	25	5	314	19	na	A
E215_4	274	4	96	1	24	8	6	2332	A
E218	4	26	2	25	5	5	19	38	D
E228_3	35	132	2	27	37	5	4	501	D
E230_4	10	63	7	1	14	7	7	na	B2

Table B.3 continued

E236_4	24	11	4	8	8	8	2	43	A
E238_3	10	11	4	8	8	8	2	10	A
E238_5	13	108	10	97	18	68	93	636	B2
E240_5	13	21	13	22	17	14	15	28	B2
E26	53	40	47	13	36	28	29	131	B2
E27	53	40	47	13	36	28	29	131	B2
E28	621	26	2	25	5	5	19	na	D
E29_1	6	11	4	223	8	78	2	1491	A
E33_4	109	65	5	1	9	13	14	517	B1
E34	4	26	2	25	5	5	19	38	D
E35_4	10	27	13	10	12	8	49	382	A
E43	21	35	27	6	5	8	4	106	D
E44_6	21	35	27	6	5	5	4	69	D
E5_5	19	23	18	24	21	2	16	32	D
E55_5	6	5	4	8	8	8	2	4	A
E55_6	4	26	2	25	5	5	19	38	D
E57	10	11	4	8	8	8	2	10	A
E58_3	6	11	4	8	8	78	2	1312	A
E61_3	274	4	96	1	24	8	6	2332	A
E64_6	53	40	47	13	36	28	29	131	B2
E67_5	10	11	4	1	8	429	2	5824	A
E70_3	na	1003	4	8	8	8	2	na	A
E71	13	21	13	22	17	14	15	28	B2

Table B.3 continued

E71_6	11	63	7	1	14	7	7	152	*Shigella*
E72	4	816	2	625	507	5	19	na	D
E72_3	4	26	2	25	5	5	19	38	D
E76_1	18	22	20	23	5	15	4	414	D
E84	6	617	57	45	11	7	7	na	B2
E88_3	6	11	4	8	8	78	2	1312	A
E88_4	10	7	4	8	12	8	2	746	A
E89_4	10	7	4	8	12	8	2	746	A
E92_5	4	26	2	25	5	5	19	38	D
E99_3	53	40	47	13	36	28	29	131	B2
Q106	4	26	39	25	5	31	19	115	D
Q108	6	4	7	1	11	3	56	270	B1
Q128	6	4	7	1	11	3	56	270	B1
Q132	24	83	4	8	8	8	2	na	A
Q142	13	108	10	97	18	68	93	636	B2
Q145	10	11	4	8	8	8	2	10	A
Q147	10	11	4	8	8	8	2	10	A
Q16_1	6	4	4	16	24	8	14	58	B1
Q186	6	212	4	1	9	48	7	1136	B1
Q196	21	35	61	52	5	5	4	394	D
Q199	6	11	4	8	8	78	2	1312	A
Q21_2	83	627	245	404	91	17	181	5148	D
Q212	10	11	4	8	8	8	2	10	A

Table B.3 continued

Q223	13	108	10	97	18	68	93	636	B2
Q23_1	1	4	4	26	20	8	381	na	B2
Q233	109	520	5	1	9	13	14	517	B1
Q240	10	11	4	8	8	8	2	10	A
Q243	21	35	61	52	5	5	4	394	D
Q245_2	6	11	4	10	7	8	6	93	A
Q249	109	65	5	1	9	13	14	517	B1
Q250_1	18	22	20	23	5	15	4	414	D
Q253	6	4	5	26	20	8	14	40	B1
Q27_1	109	65	5	1	9	13	14	517	B1
Q270	6	4	4	85	43	12	7	327	B1
Q275	109	65	5	1	9	13	14	517	B1
Q279_1	6	4	7	1	11	3	56	270	B1
Q282	536	11	22	16	522	7	2	na	A
Q284	6	4	12	476	9	7	7	4550	B2
Q288	6	11	4	223	8	78	2	1491	A
Q289	10	11	4	8	8	8	2	10	A
Q294	6	5	4	8	8	8	2	4	A
Q295	53	40	47	13	36	28	29	131	B2
Q300	10	11	4	8	8	8	2	10	A
Q307_1	10	11	4	8	8	8	2	10	A
Q308	6	11	4	10	7	8	6	93	A
Q31_6	6	5	4	323	8	8	2	na	B2

Table B.3 continued

Q310	13	108	10	97	18	68	93	636	B2
Q312_1	6	5	4	323	8	8	2	na	A
Q33_6	11	63	7	1	14	7	7	152	*Shigella*
Q34_6	11	63	7	1	14	7	7	152	*Shigella*
Q35_1	10	11	4	8	8	8	2	10	A
Q37_4	6	4	12	16	9	7	7	29	B1
Q38_1	6	5	4	8	8	8	2	4	A
Q40	10	902	4	8	8	8	2	na	A
Q49	6	29	32	16	11	8	44	4407	B2
Q51	53	40	47	13	36	28	29	131	B2
Q53	6	5	553	52	42	8	216	na	A
Q56	53	40	47	13	36	28	29	131	B2
Q65	49	185	60	45	12	35	381	na	B2
Q71_1	6	11	4	223	8	78	2	1491	A
Q85_1	10	11	4	8	8	8	2	10	A
Q86	10	11	4	1	8	8	2	34	A
Q87	6	4	12	16	24	8	14	616	B1
Q89	10	899	4	8	12	8	2	na	A
Q91	49	4	44	9	11	35	7	120	B1
R104_2	78	27	5	10	12	8	2	301	A
R109_1	603	4	573	16	9	7	12	na	B2
R110_5	6	165	4	10	7	8	6	773	A
R113_1	4	26	2	25	5	5	19	38	D

Table B.3 continued

R113_3	9	6	33	131	24	108	7	749	B1
R116_2	6	11	4	223	8	78	2	1491	A
R119_3	6	5	4	323	8	8	2	na	A
R12_1	6	4	32	16	12	8	7	164	B1
R122_4	8	11	4	8	7	8	6	181	A
R127_3	76	24	244	89	17	14	79	4554	B2
R137_1	10	27	5	8	8	7	2	226	A
R138_1	10	11	57	8	7	18	6	216	A
R17_2	6	5	4	8	8	8	2	4	A
R36_1	10	11	4	8	8	8	2	10	A
R42_2	10	11	4	8	8	8	2	10	A
R43_2	78	27	5	10	12	8	2	301	A
R46_3	10	27	5	10	96	8	2	4121	A
R55_1	6	4	12	16	24	8	14	616	B1
R56_3	18	22	17	6	5	5	4	31	D
R58_1	10	11	4	1	8	9	2	227	A
R60_2	10	11	4	8	8	8	2	10	A
R66_4	6	5	4	8	8	8	2	4	A
R66_5	6	5	4	8	8	8	2	4	A
R67_3	10	11	4	8	8	8	2	10	A
R8_4	6	11	4	8	8	78	2	1312	A
R83_3	6	4	12	1	9	2	7	295	B1
R84_2	10	11	4	361	8	8	2	3281	A

Table B.3 continued

R85_2	10	11	4	8	8	8	2	10	A
R86_1	53	40	47	13	36	28	29	131	B2
R92_5	6	4	5	26	7	8	14	200	B1

APPENDIX C

SUPPLEMENTARY TABLES FOR CHAPTER 5

Table C.1. Metagenomic yield, human read content and read quality of diarrhea and control samples used in this study

Index	Sample	case/ctl	*E. coli* pathotype	Total PE raw1	# after human cleaning	% after human cleaning	# after QC	% after QC	Final lib size
1	B001	case	DAEC	6,531,793	1,488,444	22.8	1,132,843	17.3	169M
2	B24	case	DAEC	13,156,535	12952566	98.4	10,623,704	80.7	1.5G
3	B274	case	DAEC	7,427,929	7290205	98.1	5,842,124	78.7	872M
4	B89	case	DAEC	10,171,921	1089781	10.7	908,926	8.9	126M
5	E124	case	DAEC	10,843,598	10838811	100.0	9,169,045	84.6	1.4G
6	E158	case	DAEC	8,418,618	7154333	85.0	5,874,743	69.8	878M
7	E230	case	DAEC	8,794,848	8790739	100.0	7,276,368	82.7	1.1G
8	E26	case	DAEC	5,327,047	5325794	100.0	4,079,691	76.6	609M
9	E27	case	DAEC	8,600,912	7489003	87.1	6,232,480	72.5	933M
10	Q196	case	DAEC	10,565,635	1114583	10.5	846,647	8.0	123M
11	Q308	case	DAEC	10,107,007	10083099	99.8	8,578,875	84.9	1.3G
12	Q310	case	DAEC	6,927,136	6914116	99.8	5,376,592	77.6	805M
13	Q49	case	DAEC	7,273,409	7270208	100.0	5,746,337	79.0	855M
14	Q51	case	DAEC	10,549,942	948390	9.0	766,198	7.3	113M
15	Q56	case	DAEC	10,155,142	7471039	73.6	6,260,221	61.6	916M
16	Q65	case	DAEC	8,922,179	6087254	68.2	5,074,850	56.9	759M
17	B228	case	EPEC	9,257,000	9220639	100	7,785,343	84.1	1.2G
18	B56	case	EPEC	10,822,204	9224643	85	7,593,577	70.2	1.1G
19	B69	case	EPEC	7,542,022	7518431	100	6,113,656	81.1	909M
20	E162	case	EPEC	6,938,905	4948559	71	3,997,479	57.6	598M
21	E187	case	EPEC	9,032,039	8537949	95	7,158,475	79.3	1.1G
22	E205	case	EPEC	7,036,190	7034607	100	5,384,029	76.5	811M
23	Q233	case	EPEC	8,942,343	8941264	100	7,252,948	81.1	1.1G
24	Q300	case	EPEC	9,834,934	9772091	99	8,233,965	83.7	1.2G
25	R126	case	EPEC	11,185,025	404832	4	316,760	2.8	47M
26	R135	case	EPEC	10,490,440	4523998	43	3,866,783	36.9	581M
27	B109	case	ETEC	9,192,633	9189530	100	7675507	83.5	1.2G
28	B200	case	ETEC	12,509,793	12494918	100	10313659	82.4	1.5G
29	B244	case	ETEC	8,295,416	7134222	86	5775817	69.6	862M

Table C.1 continued

30	B255	case	ETEC	9,856,355	9854564	100	8432852	85.6	1.3G
31	B295	case	ETEC	7,378,814	7376611	100	5930153	80.4	888M
32	B45	case	ETEC	11,491,641	11487903	100	9823847	85.5	1.4G
33	B62	case	ETEC	8,850,080	8845711	100	7358733	83.1	1.1G
34	B64	case	ETEC	8,999,675	8990531	100	7637068	84.9	1.1G
35	B68	case	ETEC	8,218,792	7986605	97	6509243	79.2	964M
36	E184	case	ETEC	6,314,505	6287420	100	5175784	82.0	771M
37	Q294	case	ETEC	13,322,967	13312996	100	11105678	83.4	1.7G
38	Q53	case	ETEC	1,286	1,213	94.32	388	32	49K
39	Q105	control	NEG	5,377,446	5,375,319	99.96	4016518	74.69	1.2G
40	Q127	control	NEG	7,778,804	7,776,111	99.97	6304665	81.05	1.9G
41	Q158	control	NEG	7502883	7501382	99.98	5800986	77.32	1.7G
42	Q116	control	NEG	8405358	8402475	99.97	6737952	80.16	2G
43	Q157	control	NEG	6500061	6495920	99.94	5136050	79.02	1.5G
44	Q101	control	NEG	9119586	9117396	99.98	7822285	85.77	2.3G
45	Q131	control	NEG	7899998	7897301	99.97	6472827	81.93	1.9G
46	R15	control	NEG	8506033	8442959	99.26	6998463	82.28	2.1G
47	R22	control	NEG	7069379	7066763	99.96	5426039	76.75	1.6G
48	R26	control	NEG	7603692	7601120	99.97	6394649	84.10	1.9G
49	R80	control	NEG	7055082	7052399	99.96	5731005	81.23	1.7G
50	R81	control	NEG	6062506	6060167	99.96	5026868	82.92	1.5G
51	R91	control	NEG	9315213	9313429	99.98	7569775	81.26	2.2G
52	R105	control	NEG	9295732	9292607	99.97	7799973	83.91	2.3G
53	R130	control	NEG	9161454	9161060	99.99	7756239	84.66	2.3G
54	R134	control	NEG	9271545	9270609	99.99	7793289	84.06	2.3G
55	R124	control	NEG	8928121	8925587	99.97	7258528	81.30	2.1G
56	R129	control	NEG	7789512	7788783	99.99	6659108	85.49	2G
57	R131	control	NEG	7470866	7459793	99.85	6145847	82.26	1.8G
58	R25	control	NEG	10159250	10157817	99.99	8966203	88.26	2.7G
59	R29	control	NEG	7547198	7546293	99.99	6386418	84.62	1.9G
60	R40	control	NEG	8606454	8606023	99.99	7255722	84.31	2.2G
61	R97	control	NEG	10513556	10486995	99.75	8803096	83.73	2.6G

Table C.2. General statistics of *E. coli* isolate genomes and MAGs recovered from diarrheal metagenomes.

Sample_ID	Pathotype	Type	Size (Mb)	Contigs	GC %	N50	Coverage	Completeness %	Contamination %
Q56_bin2	DAEC	MAG	4.9	113	50.7	146837	82.4X	99.59	1.27
E230_bin3	DAEC	MAG	5.2	149	50.5	131672	90.9X	99.12	1.01
Q196_bin1	DAEC	MAG	4.9	541	51.8	20774	23.4X	97.1	1.85
E124_bin9	DAEC	MAG	4.7	244	51.3	34664	16.7X	96.91	0.99
Q51_bin1	DAEC	MAG	4.1	1516	50.4	1516	25X	70.3	3.6
Q65_bin3	DAEC	MAG	4.6	286	50.8	36969	76.7X	96.24	2.78
E158_bin1	DAEC	MAG	3.7	861	51.9	6609	123X	76.23	2.06
R135_bin8	EPEC	MAG	4.9	961	51.1	8631	11X	95.21	2.84
B45_bin12	ETEC	MAG	4.6	179	50.9	52598	18.6X	98.84	0.55
B295_bin5	ETEC	MAG	5.1	465	51.2	26676	12.3X	97.99	3.78
E184_bin4	ETEC	MAG	4.7	530	51.1	16705	32.3X	94.37	2.6
B62_bin13	ETEC	MAG	4.3	1343	51.8	2835	6.3X	70.2	5
B200_bin15	ETEC	MAG	4.4	720	51.3	10164	12.X	93.24	1.98
Q294_bin5	ETEC	MAG	4.3	611	51.1	13230	71X	90.93	2.48
B001_5	DAEC	Isolate	5.1	257	50.5	61824	32X	99.1	0.99
B24_1	DAEC	Isolate	4.8	253	49.7	55219	162X	98.6	1.06
B274_2	DAEC	Isolate	4.9	254	50	59071	335X	98.87	0.63
B89_1	DAEC	Isolate	4.8	194	50.6	68348	89X	98.35	0.42
E124_5	DAEC	Isolate	5.1	204	50.53	86732	36X	99.47	1.1
E158	DAEC	Isolate	5.2	202	50.59	77037	27X	99.56	0.84
E230_4	DAEC	Isolate	5.8	1630	50	5050	40X	91.89	8.29
E26	DAEC	Isolate	5.3	206	50.44	105975	28X	99.45	1.62
E27	DAEC	Isolate	5.1	142	50.2	103201	87X	99.59	0.92
Q196	DAEC	Isolate	5.1	309	50.42	34191	17X	99.24	0.59
Q308	DAEC	Isolate	4.9	192	50.66	66199	31X	99.2	0.74
Q310	DAEC	Isolate	5	205	50.54	59022	27X	99.17	0.78
Q49	DAEC	Isolate	5	173	50.59	111283	28X	99.43	0.48
Q51	DAEC	Isolate	5.2	252	50.52	55381	17X	99.51	1.65
Q56	DAEC	Isolate	5	137	50.63	118040	22X	99.5	1.1
Q65	DAEC	Isolate	5.1	1223	50.07	6466	28X	92.74	4.6
B228_2	EPEC	Isolate	5.1	192	49.7	80361	428X	99.47	0.69
B56	EPEC	Isolate	na	na	na	na	na	na	na
B69_1	EPEC	Isolate	4.8	168	449.6	66315	78X	98.82	0.23
E162	EPEC	Isolate	4.7	148	50.65	89603	23X	99.47	0.39
E187	EPEC	Isolate	5	298	50.3	42003	38X	98.52	0.57
E205	EPEC	Isolate	4.5	131	50.58	64835	27X	99.17	0.3
Q233	EPEC	Isolate	4.8	625	50.42	13401	12X	97.27	1.46
Q300	EPEC	Isolate	4.5	131	49.6	64835	64X	99.17	0.3
R126	EPEC	Isolate	na	na	na	na	na	na	na
B109_1	ETEC	Isolate	4.7	403	50.5	22016	17X	99.19	0.73
B200_2	ETEC	Isolate	4.8	245	50.63	42883	19X	99.34	0.37
B244_3	ETEC	Isolate	4.8	201	48.3	66123	289X	99.34	0.55
B255_1	ETEC	Isolate	4.7	246	48.7	49462	167X	98.64	0.39
B295_2	ETEC	Isolate	4.7	255	48.6	49462	118X	98.37	0.39
B45	ETEC	Isolate	4.8	318	50.6	30625	26X	98.34	0.66
B62_5	ETEC	Isolate	4.8	274	50.5	43727	140X	98.96	0.55
B68_1	ETEC	Isolate	4.7	200	49.5	64686	352X	99.38	0.74
E184_3	ETEC	Isolate	5	182	49.7	88848	46X	99.52	1.29
Q294	ETEC	Isolate	4.8	215	50.52	53151	25X	98.64	0.56
Q53	ETEC	Isolate	5.2	1611	50.1	43.03	18X	88.57	7.63

Table C.3. Epidemiology of DAEC isolates assigned to clonal complexes based on MLSA and core phylogeny phylogroups. Strain ID in bold denotes those isolates specific for each pathotype group. Isolates for which no sequence type (ST) could be assigned, are denoted with na.

Strain ID	Pathotype	case/control	ST	Phylogroup	Clonal Complex
E124_6	DAEC	case	1193	B2	1
E238_5	DAEC	control	636	B2	1
Q223	DAEC	case	636	B2	1
Q310	DAEC	case	636	B2	1
Q142	DAEC	case	636	B2	1
Q56	DAEC	case	131	B2	2
E27	DAEC	case	131	B2	2
R86	DAEC	case	131	B2	2
E170	DAEC	case	131	B2	2
E179_6	DAEC	control	na	B2	2
E26	DAEC	case	131	B2	2
Q51	DAEC	case	131	B2	2
E205	EPEC	case	394	D	3
B228_2	EPEC	case	394	D	3
Q196	DAEC	case	394	D	3
Q243	DAEC	case	394	D	3
E173	DAEC	case	394	D	3
B147_1	DAEC	control	394	D	3
E230_4	DAEC	case	na	B1	4
E144	EPEC	case	40	B1	4
Q253	EPEC	control	40	B1	4
E124_5	DAEC	case	200	B1	4
E106_1	EAEC	case	200	B1	4
B108_3	DAEC	control	200	B1	4
R92_5	EAEC	control	200	B1	4
Q65	DAEC	case	na	B1	5
Q23_1	DAEC	case	na	B1	5
Q91	DAEC	case	120	B1	5
E158	DAEC	case	38	D	6
E100_4	DAEC	case	38	D	6
E212_5	DAEC	case	38	D	6
E16_6	DAEC	control	38	D	6
E92_5	DAEC	control	38	D	6
B75_5	DAEC	case	38	D	6
C14_2A	DAEC	case	38	D	6

Table C.4. Epidemiology of ETEC isolates assigned to clonal complexes based on MLSA and core phylogeny phylogroups. Strain ID in bold denotes those isolates specific for each pathotype group. Isolates for which no sequence type (ST) could be assigned, are denoted with na.

Strain ID	Pathotype	case/control	ST	Phylogroup	Clonal Complex
B244_3	ETEC	case	4	A	1
Q294	ETEC	case	4	A	1
Q38_1	ETEC	case	4	A	1
R119_3	ETEC	control	na	A	1
B295_2	ETEC	case	na	A	1
B255_1	ETEC	case	na	A	1
B46_1	ETEC	control	4	A	1
B45_2	ETEC	case	4	A	1
R66_4	ETEC	case	4	A	1
R66_5	ETEC	case	4	A	1
R17	ETEC	case	4	A	1
C46_4	ETEC	control	4	A	1
B62_5	ETEC	case	4	A	1
E195_3	ETEC	case	94	B1	2
C85_1	ETEC	case	94	B1	2
C86_4	EAEC	control	na	B1	2
R113_3	ETEC	case	749	B1	2
B109_1	ETEC	case	3857	B1	2
B84_2	ETEC	case	3857	B1	2
B68_1	ETEC	case	3857	B1	2
B200_2	ETEC	case	641	B1	2
E61_3	ETEC	case	2332	A	3
E184_3	ETEC	case	2332	A	3
E215_4	ETEC	control	2332	A	3

Table C.5. Epidemiology of EPEC isolates assigned to clonal complexes based on MLSA and core phylogeny phylogroups. Strain ID in bold denotes those isolates specific for each pathotype group.

Strain ID	Pathotype	case/control	ST	Phylogroup	Clonal Complex
E205	EPEC	case	394	D	1
B228_2	EPEC	case	394	D	1
Q196	DAEC	case	394	D	1
Q243	DAEC	case	394	D	1
E173	DAEC	case	394	D	1
B147_1	DAEC	control	394	D	1
E240_5	EPEC	control	28	B2	2
E71	EPEC	case	28	B2	2
E187_2	EPEC	case	28	B2	2
B235_3	EPEC	case	28	B2	2
B231_1	EPEC	control	28	B2	2
Q275	EPEC	control	517	B1	3
C26_2	EPEC	control	517	B1	3
Q233	EPEC	case	517	B1	3
Q249	EPEC	control	517	B1	3
E33_4	EPEC	control	517	B1	3
B259_1	EPEC	control	517	B1	3
Q27_1	EPEC	control	517	B1	3
Q300	EPEC	case	10	A	4
E20_1	EAEC	control	10	A	4
C21_2	EAEC	control	10	A	4
C21_4	EAEC	control	10	A	4